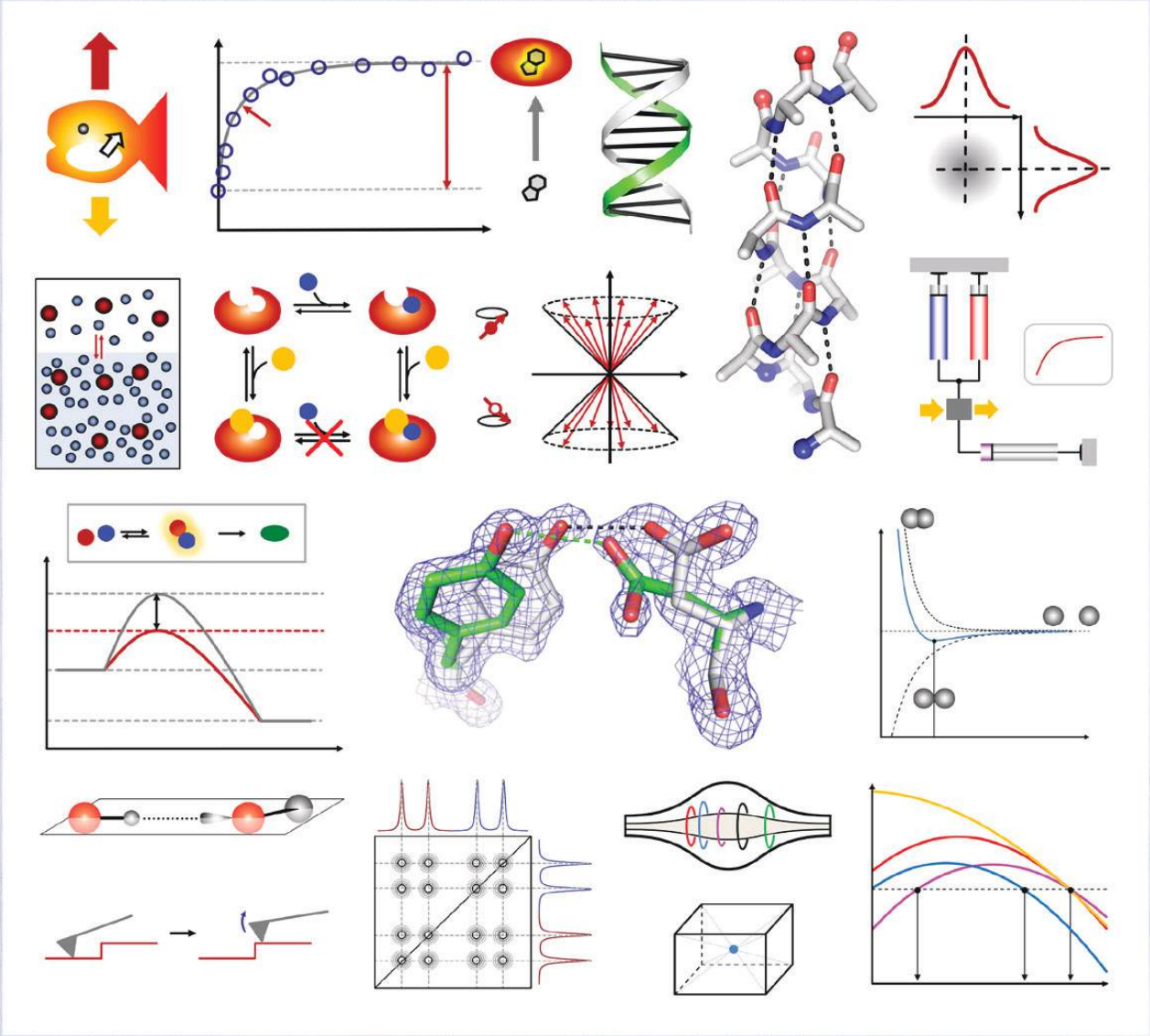
SECOND EDITION

Biophysical Chemistry



Dagmar Klostermeier and Markus G. Rudolph

CRC Press
Taylor & Francis Group

Biophysical Chemistry

Biophysical Chemistry explores the concepts of physical chemistry and molecular structure that underlie biochemical processes. Ideally suited for undergraduate students and scientists with backgrounds in physics, chemistry, or biology, it is also equally accessible to students and scientists in related fields as the book concisely describes the fundamental aspects of biophysical chemistry and puts them into a biochemical context.

This second edition has been fully updated throughout with novel techniques, with a new chapter on advances in cryo-electron microscopy and exciting new content throughout on big data techniques, structural bioinformatics, systems biology and interaction networks, and artificial intelligence and machine learning.

The book is organized in four parts, covering thermodynamics, kinetics, molecular structure and stability, and biophysical methods. Cross-references within and between these parts emphasize common themes and highlight recurrent principles. End of chapter problems illustrate the main points explored and their relevance for biochemistry, enabling students to apply their knowledge and to transfer it to laboratory projects.

Key Features:

- Connects principles of physical chemistry to biochemistry.
- Emphasizes the role of organic reactions as tools for modification and manipulation of biomolecules.
- Includes a comprehensive section on the theory of modern biophysical methods and their applications.

Dagmar Klostermeier is of the University of Muenster, Germany.

Markus G. Rudolph is of the Pharma Research & Early Development Division in the Roche Group, Switzerland.

Biophysical Chemistry

Second Edition

Dagmar Klostermeier • Markus G. Rudolph

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

Designed cover image: Dagmar Klostermeier and Markus G. Rudolph

Second edition published 2025
by CRC Press
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

© 2025 Dagmar Klostermeier and Markus G. Rudolph

First edition published by CRC Press 2017

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC please contact mpkbookspermissions@tandf.co.uk

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Names: Klostermeier, Dagmar, author. | Rudolph, Markus G., author.
Title: Biophysical chemistry / Dagmar Klostermeier and Markus G. Rudolph.
Description: 2nd edition. | Boca Raton, FL : CRC Press, 2024. | Includes bibliographical references and index. | Summary: "Biophysical Chemistry explores the concepts of physical chemistry and molecular structure that underlie biochemical processes. Ideally suited for undergraduate students and scientists with backgrounds in physics, chemistry or biology, it is also equally accessible to students and scientists in related fields as the book concisely describes the fundamental aspects of biophysical chemistry, and puts them into a biochemical context. This second edition has been fully updated throughout with novel techniques, with a new chapter on advances in cryo-electron microscopy and exciting new content throughout on big data techniques, structural bioinformatics, systems biology and interaction networks, and artificial intelligence and machine learning. The book is organized in four parts, covering thermodynamics, kinetics, molecular structure and stability, and biophysical methods. Cross-references within and between these parts emphasize common themes and highlight recurrent principles. End of chapter problems illustrate the main points explored and their relevance for biochemistry, enabling students to apply their knowledge and to transfer it to laboratory projects"-- Provided by publisher.
Identifiers: LCCN 2024001522 | ISBN 9781032060835 (hardback) | ISBN 9781032068152 (paperback) | ISBN 9781003203988 (ebook)
Subjects: LCSH: Physical biochemistry. | Thermodynamics. | Molecular structure.
Classification: LCC QD476.2 .K56 2024 | DDC 572/.43--dc23/eng/20240617
LC record available at https://lccn.loc.gov/2024001522

ISBN: 9781032060835 (hbk)
ISBN: 9781032068152 (pbk)
ISBN: 9781003203988 (ebk)

DOI: 10.1201/9781003203988

Typeset in Times
by Deanta Global Publishing Services, Chennai, India

Contents

Preface to the first edition xv
Preface to the second edition xvii
Acknowledgments to the first edition xix
Acknowledgments to the second edition xxi
Authors xxii

PART I THERMODYNAMICS

Chapter 1 Systems, state functions, and the laws of thermodynamics 3

1.1 Systems and their surroundings 3
1.2 State variables and state functions 4
1.3 The internal energy *U* and the first law of thermodynamics 6
1.3.1 Internal energy, heat, and work 6
1.3.2 The first law of thermodynamics 7
1.3.3 The ideal gas: a convenient system to understand thermodynamic principles 8
1.3.4 Changes in the state of an ideal gas 11
1.3.5 Thermodynamic cycles: back and forth or round and round 20
1.3.6 The temperature dependence of the internal energy *U* 24
1.4 The enthalpy *H* 25
1.5 The entropy *S* and the second law of thermodynamics 28
1.5.1 Predicting spontaneity of processes: dissipation of heat and matter 28
1.5.2 Entropy and heat 29
1.5.3 Temperature dependence of the entropy 32
1.5.4 The third law of thermodynamics and absolute entropy 33
1.5.5 Entropy and order: the statistic interpretation 34
1.6 The free energy *G*: combining system and surroundings 37
1.6.1 Entropy- and enthalpy-driven reactions 39
1.6.2 Pressure and temperature dependence of the free energy 41
1.6.3 Standard states 43
1.6.4 Relation of free energy, enthalpy, and entropy to molecular properties 44
1.7 The chemical potential μ 46
1.7.1 The chemical potential as a driving force for chemical reactions 47
1.7.2 The chemical potential and stable states: phase diagrams 48

1.7.3 Pressure and temperature dependence of the chemical potential 51
1.7.4 The chemical potential as a partial molar property 51
1.7.5 The chemical potential of compounds in mixtures 53
1.7.6 The chemical potential of solutions 55
1.7.7 Colligative properties 57
Questions 64
References 67

Chapter 2 Energetics and chemical equilibria 68

2.1 The free energy change and the equilibrium constant 68
2.1.1 Temperature dependence of the equilibrium constant 70
2.1.2 The principle of Le Chatelier 71
2.2 Binding and dissociation equilibria and affinity 72
2.3 Protolysis equilibria: the dissociation of acids and bases in water 75
2.4 Thermodynamic cycles, linked functions, and apparent equilibrium constants 76
Questions 84
References 86

Chapter 3 Statistical thermodynamics 87

3.1 Probabilities, statistical weights, energy, and the Boltzmann distribution 87
3.2 Sum of states: the partition function 90
3.2.1 General form of the partition function 90
3.2.2 The partition function for systems with degenerate states 92
3.2.3 From the molecular partition function to the partition function of ensembles 93
3.3 Statistical thermodynamics and structural changes in biomolecules 95
3.3.1 General description of structural transitions 95
3.3.2 Applications of the zipper model to structural transitions of biomolecules 99
3.4 Statistical thermodynamics and binding equilibria 100
3.4.1 Multiple equivalent, non-interacting binding sites 101
3.4.2 Multiple classes of non-equivalent, non-interacting binding sites 104
3.4.3 Cooperative binding sites 105
Questions 107
References 109

Chapter 4 Thermodynamics of transport processes 110

4.1 Diffusion 110
4.2 The chemiosmotic hypothesis 116
4.3 Active and passive transport 118
4.4 Directed movement by the Brownian ratchet 120
Questions 124
References 125

Chapter 5 Electrochemistry 126

5.1 Redox reactions and electrochemical cells 126
5.2 Types of half-cells 129
5.3 Standard electrode potentials 130
5.4 The Nernst equation 131
5.5 Measuring pH values 133
5.6 Redox reactions in biology 134
5.6.1 Cysteines and disulfide bonds 134
5.6.2 The respiratory chain 134
5.6.3 The light reaction in photosynthesis 135
5.7 The electrochemical potential and membrane potentials 136

5.8 Electrophysiology: patch-clamp methods to measure ion flux through ion channels . . 139
Questions 141
References 142

PART II KINETICS

Chapter 6 Reaction velocities and rate laws 145

Questions 149

Chapter 7 Integrated rate laws for uni- and bimolecular reactions 150

Questions 161

Chapter 8 Reaction types. 162

8.1 Reversible reactions 162
8.2 Parallel reactions 166
8.3 Consecutive reactions 168
Questions 174
References 174

Chapter 9 Deriving integrated rate laws by solving sets of differential equations with matrix algebra . . . 175

9.1 Reversible reactions 175
9.2 Consecutive, irreversible reactions 180
9.3 Reversible reaction followed by an irreversible second step 183
Questions 188
References 188

Chapter 10 Rate-limiting steps 189

Questions 193
References 193

Chapter 11 Binding reactions: One-step and two-step binding 194

Questions 202
References 202

Chapter 12 Single-molecule kinetics. 203

12.1 Integrated rate laws, dwell times, dwell-time and cumulative dwell-time distributions 203
12.2 Single-molecule kinetics and the principle of ergodicity 204
12.3 Single-molecule kinetics of reversible reactions 206
12.4 The limits of dwell-time histograms: potential pitfalls 209
12.5 Double-exponential dwell-time histograms 210
Questions 212
References 212

Chapter 13 Steady-state (enzyme) kinetics 213

13.1 Rapid equilibrium (Michaelis-Menten formalism) 214
13.2 Steady-state approximation (Briggs-Haldane formalism) 216
13.3 pH dependence 220
13.4 Two or more non-interacting active sites 224
13.5 Two or more interacting active sites: cooperativity and the Hill equation 228
13.6 Inhibition of enzyme activity 232
13.6.1 Product inhibition in reversible reactions 232
13.6.2 Competitive inhibition 235

13.6.3 Non-competitive inhibition 236
13.6.4 Mixed inhibition 238
13.6.5 Quantifying enzyme inhibition: inhibition constants *versus* IC_{50} values 240
13.6.6 Time-dependent inhibition 242
Questions 251
References 254

Chapter 14 Complex reaction schemes and their analysis 255

14.1 Binding of two substrates 255
14.1.1 Random binding 255
14.1.2 Ordered binding 257
14.2 Ping-pong mechanism 260
14.3 Net rate constants and transit times 261
Questions 264
References 265

Chapter 15 Temperature dependence of rate constants 266

15.1 The Arrhenius equation 266
15.2 Transition state theory 266
15.3 Collision theory 269
15.4 Kinetic and thermodynamic control of reactions 270
Questions 271

Chapter 16 Principles of catalysis 273

16.1 Heterogeneous catalysis: the Langmuir-Hinshelwood formalism for surface-catalyzed reactions 274
16.2 Homogeneous catalysis by enzymes 276
16.2.1 Acid-base catalysis 278
16.2.2 Electrostatic and covalent catalysis 281
16.2.3 Intramolecular catalysis and effective concentrations 282
Questions 283
References 283

PART III MOLECULAR STRUCTURE AND STABILITY

Chapter 17 Molecular structure and interactions 287

17.1 Isomers: configuration and conformation 287
17.1.1 Configurations 288
17.1.2 Conformations 290
17.2 The configurational and conformational energy of molecules 290
17.2.1 Covalent interactions 290
17.2.2 Non-covalent interactions 295
17.3 Energy landscapes and the most stable conformation 305
Questions 306
References 307

Chapter 18 Proteins 308

18.1 Amino acids and the peptide bond 308
18.1.1 Properties of the twenty canonical amino acids 308
18.1.2 The peptide bond 311
18.1.3 Side-chain rotamers 313

18.1.4 Post-translational modifications....313
18.1.5 Expansion of the genetic code....320
18.2 Protein sequence....322
18.2.1 Analysis of individual sequences for global and local properties....323
18.2.2 Sequence alignment....325
18.3 Protein structure....327
18.3.1 Helical secondary structure elements....329
18.3.2 β-strands and their super-secondary structures (β-sheets)....334
18.3.3 Reverse turns....337
18.3.4 Protein domains and tertiary structure....341
18.3.5 Quaternary structure....344
18.3.6 Protein-protein interfaces....347
18.3.7 Protein-ligand interactions....351
18.4 Membrane proteins and their lipid environment....354
18.4.1 Biological roles of lipids and membranes....354
18.4.2 Types of lipids....355
18.4.3 Super-structures formed by amphiphiles....356
18.4.4 Properties and structure of membrane proteins....364
18.4.5 Solubilization of membrane proteins....367
18.5 Folding and stability....369
18.5.1 Driving forces for protein folding....369
18.5.2 First folding experiments and the Levinthal paradox....372
18.5.3 Energy landscapes for protein folding....373
18.5.4 Mathematical description of the two-state model....374
18.5.5 Folding pathways and mechanisms of protein folding....380
18.5.6 Protein folding diseases....388
Questions....389
References....392

Chapter 19 Nucleic acids....395

19.1 Nucleobases, nucleosides, and nucleotides....396
19.1.1 Hydrogen bonds between nucleobases – base pairs....397
19.1.2 Non-standard nucleobases in DNA....399
19.1.3 Non-standard nucleobases in RNA....400
19.2 Ribose and nucleobase conformations....401
19.2.1 Sugar pucker....401
19.2.2 Syn- and anti- conformations....402
19.3 Primary structure of nucleic acids....403
19.3.1 Nucleic acid sequences and codon usage....403
19.3.2 Backbone conformation....405
19.4 Base-pair geometry and stacking....406
19.4.1 Importance of base-pair stacking for double helix formation....407
19.4.2 Base-pair geometries....408
19.5 DNA structures and conformations....409
19.5.1 DNA double helical structures....409
19.5.2 Triple and quadruple DNA helices....412
19.5.3 Higher-order DNA structures....414
19.5.4 DNA interactions with proteins and ligands....422
19.6 RNA structure....428
19.6.1 RNA secondary structure and their prediction....429
19.6.2 RNA tertiary structure....431
19.6.3 Ribozymes, riboswitches, and small molecule binding to RNA....434

19.6.4 RNA folding 436
19.6.5 Ribonucleoprotein (RNP) complexes 437
Questions 439
References 441

Chapter 20 Macromolecular modeling 443

20.1 Molecular mechanics and dynamics 443
20.1.1 Force fields 444
20.1.2 Energy minimization 445
20.1.3 Simulated annealing 447
20.1.4 Coarse-grained modeling 448
20.1.5 Molecular mechanics and dynamics 449
20.2 Principles of neural networks and machine learning 454
20.3 Simulating protein folding 456
20.3.1 Homology modeling 456
20.3.2 Sequence-based structure prediction 458
20.3.3 Quality assessment of predicted three-dimensional protein models 461
20.3.4 Biophysical value and limitations of predicted structural models 463
Questions 465
References 466
Online resources 466

PART IV METHODS

Chapter 21 Optical spectroscopy 471

21.1 Interaction of light and matter 471
21.1.1 Light as an electromagnetic wave 471
21.1.2 Principles of spectroscopy: transitions in two-state systems 473
21.2 Absorption 475
21.2.1 Electronic, vibronic, and rotational energy levels 475
21.2.2 Transitions and transition dipoles 476
21.2.3 The Lambert-Beer law 478
21.2.4 Solvent effects and influence of the local environment 479
21.2.5 Instrumentation 480
21.2.6 Biological chromophores 481
21.2.7 Applications 485
21.2.8 Potential pitfalls 492
21.3 Linear and circular dichroism 493
21.3.1 Linearly polarized light and linear dichroism 493
21.3.2 Circularly polarized light and circular dichroism 496
21.3.3 Instrumentation 500
21.3.4 Biological chromophores that show circular dichroism 501
21.3.5 Applications 502
21.3.6 Potential pitfalls 503
21.4 Infrared spectroscopy 504
21.4.1 Bond vibrations: the harmonic oscillator 504
21.4.2 Molecule geometry, degrees of freedom, and vibrational modes 506
21.4.3 Instrumentation 508
21.4.4 Applications 509
21.5 Fluorescence 510
21.5.1 General considerations 510
21.5.2 Instrumentation 512

21.5.3 Quantum yield and lifetime 513
21.5.4 Fluorophores and fluorescent labeling 514
21.5.5 Applications 523
21.5.6 Potential pitfalls 531
21.5.7 Fluorescence quenching 535
21.5.8 Fluorescence anisotropy 537
21.5.9 Time-resolved fluorescence 542
21.5.10 Förster Resonance Energy Transfer 550
Questions 561
References 565

Chapter 22 Magnetic resonance 570

22.1 Nuclear magnetic resonance 570
22.1.1 Nuclear spins and the Zeeman effect 570
22.1.2 A one-dimensional NMR spectrum: Larmor frequency, chemical shift, *J*-coupling, and multiplicity 572
22.1.3 The nuclear Overhauser effect: distance information 577
22.1.4 Magnetization and its relaxation to equilibrium: Fourier transform-NMR and the free induction decay 580
22.1.5 Two-dimensional FT-NMR: COSY and NOESY 585
22.1.6 Extending NMR to structure determination of large molecules 592
22.1.7 NMR and dynamics 594
22.1.8 Solid-state NMR and biology 597
22.1.9 NMR and imaging 598
22.2 Electron paramagnetic resonance 599
22.2.1 Principle of electron paramagnetic resonance 599
22.2.2 Spin-spin interactions: hyperfine coupling of unpaired electrons with nuclei 601
22.2.3 EPR probes and spin labeling 601
22.2.4 EPR as a probe for mobility and dynamics 603
22.2.5 EPR as a probe for accessibility 605
22.2.6 Measuring spin-spin distances 606
22.2.7 Distance determination by pulsed EPR: PELDOR/DEER 606
Questions 609
References 611

Chapter 23 Solution scattering 615

23.1 Light scattering 615
23.1.1 Static light scattering 615
23.1.2 Dynamic light scattering 620
23.1.3 Raman scattering 622
23.2 Small angle scattering 624
23.2.1 Scattering of X-rays and neutrons 624
23.2.2 SAS intensity distribution 627
23.2.3 Distance distribution function 631
23.2.4 Small angle X-ray scattering 632
23.2.5 Small angle neutron scattering 635
Questions 637
References 638

Chapter 24 Crystallography 640

24.1 Phase problem and requirement for crystals 640
24.2 Crystallization of macromolecules 641
24.3 Symmetry and space groups 647

24.4 X-ray crystallography 652
21.4.1 Generation of X-rays 653
24.4.2 X-ray diffraction from crystals 657
24.4.3 Diffraction data collection and analysis 661
24.4.4 Phasing methods 665
24.4.5 Electron density and model building 673
24.4.6 Model refinement and validation 675
24.5 Time-resolved crystallography 679
24.6 Serial crystallography and X-ray free electron lasers 680
24.7 Neutron crystallography 682
24.8 Electron crystallography 682
Questions 683
References 685
Online resources 686

Chapter 25 Fluorescence imaging and microscopy 687

25.1 Optical principles of microscopy 687
25.1.1 Focusing and collecting light by optical lenses 688
25.1.2 Microscopes: how to achieve magnification with optical lenses 690
25.1.3 The diffraction limit of optical resolution 691
25.2 Wide-field fluorescence microscopy 693
25.3 Confocal scanning microscopy 695
25.4 Total internal reflection microscopy 697
25.5 Fluorescence lifetime imaging microscopy 700
25.6 Fluorescence (cross-)correlation spectroscopy 701
25.6.1 Fluorescence correlation spectroscopy 701
25.6.2 FCS to monitor binding events 703
25.6.3 Fluorescence cross-correlation spectroscopy 707
25.7 Single-molecule fluorescence microscopy 708
25.7.1 Principles of single-molecule microscopy 708
25.7.2 Why single molecules? 711
25.7.3 Localization and tracking of single molecules 712
25.7.4 Kinetic information from single-molecule microscopy 714
25.7.5 Colocalization of molecules 716
25.7.6 Single-molecule FRET 718
25.8 Super-resolution microscopy 723
Questions 727
References 728

Chapter 26 Electron microscopy 731

26.1 Outline of an electron microscope 733
26.1.1 Generation of electron beams 734
26.1.2 Electron energy and resolution limit of EM 734
26.1.3 Electromagnetic lenses 736
26.1.4 Electron detection and detector noise 739
26.2 Electron-matter interactions 741
26.2.1 Elastic scattering of electrons 741
26.2.2 Inelastic scattering of electrons 742
26.2.3 Radiation damage 742
26.3 Sample preparation for EM 743
26.3.1 EM grids and sample supports 744
26.3.2 Metal shadowing for SEM 744
26.3.3 Negative staining for TEM 745

26.3.4 Native samples in vitreous ice 745
26.4 TEM imaging and diffraction 746
26.4.1 Plane waves, defocus, and phase shifts 748
26.4.2 Image contrast, amplitude contrast, and phase contrast 749
26.4.3 Contrast transfer function 751
26.5 Three-dimensional electron microscopy 756
26.5.1 Cryo-electron tomography 756
26.5.2 Single-particle cryo-EM 759
26.5.3 Resolution and interpretation of cryo-EM potential maps 762
Questions 764
References 765

Chapter 27 Scanning probe microscopy and force measurements 766

27.1 Scanning probe microscopy: scanning tunneling, scanning force, and atomic force microscopy 766
27.2 Force measurements 768
27.2.1 Effect of forces on the energetics and kinetics of macromolecules 768
27.2.2 Force spectroscopy by AFM 770
27.2.3 Optical tweezers 776
27.2.4 Magnetic tweezers 782
Questions 786
References 786

Chapter 28 Transient kinetic methods 791

28.1 Stopped flow 791
28.2 Quench flow 794
28.3 Laser flash photolysis 795
28.4 Relaxation kinetics: pressure- and temperature-jump 797
Questions 799
References 799

Chapter 29 Molecular mass, size, and shape 801

29.1 Mass spectrometry 801
29.1.1 Ionization 802
29.1.2 Ion storage and manipulation 804
29.1.3 Detection 811
29.1.4 Mass spectra 813
29.1.5 Applications 813
29.2 Analytical ultracentrifugation 820
29.2.1 Instrumentation and detection systems 820
29.2.2 Behavior of a molecule in a gravitational field 822
29.2.3 Sedimentation velocity 826
29.2.4 Sedimentation equilibrium 831
29.2.5 Zonal, band, or isopycnic centrifugation 834
29.3 Surface plasmon resonance 837
29.3.1 Physical background of SPR 838
29.3.2 Principle and information content of an SPR experiment 839
29.3.3 Mass transport limitation 841
29.3.4 Receptor immobilization on the sensor surface 842
29.3.5 Stoichiometry of binding in an SPR experiment 845
29.3.6 Specificity of binding in an SPR experiment 845
Questions 846
References 850

Online resources 852

Chapter 30 Calorimetry 853

30.1 Isothermal titration calorimetry 853
30.1.1 General principle 853
30.1.2 ITC data analysis 855
30.1.3 Origin of enthalpic changes 857
30.1.4 Practical considerations 858
30.1.5 Measuring high affinities with ITC by competition 861
30.1.6 Measuring Michaelis-Menten enzyme kinetics with ITC 861
30.2 Differential scanning calorimetry 864
30.2.1 General principle 864
30.2.2 Two-state unfolding of macromolecules 867
30.2.3 Two-state unfolding with subunit dissociation 869
Questions 870
References 872

APPENDIX

Chapter 31 Mathematical concepts used in this book 875

31.1 Sums and products 875
31.2 Quadratic equation 876
31.3 Binomial coefficients 876
31.4 Trigonometry 877
31.5 Logarithms and exponentials 878
31.6 Differentiation and integration 880
31.6.1 Implicit derivatives 881
31.6.2 Partial and total derivatives 881
31.6.3 Integration 882
31.6.4 Differential equations 883
31.7 Partial fractions 885
31.8 l'Hôpital's rule 886
31.9 Vectors 886
31.10 Matrices 888
31.10.1 Basic operations on matrices 889
31.10.2 Solving systems of linear equations by Gaussian elimination 890
31.10.3 Eigenvalues and Eigenvectors to solve systems of linearly independent equations 891
31.11 Complex numbers 893
31.12 Basic elements of statistics 895
31.13 Error propagation 897
31.14 Series expansion 899
31.14.1 Taylor series 899
31.14.2 Fourier series 900
31.15 Fourier transformation 901
31.16 Convolution 903

Index 905

Preface to the first edition

Writing this book, we wanted to present the concepts of physical chemistry and molecular structure that underlie biochemical processes from a biochemist's perspective. The content of the book is equally well accessible to students and scientists with backgrounds in physics, chemistry, or biology, but is particularly intended for bachelor and master students of life science curricula. The book will also be useful as an introductory text to more advanced students from other fields, and should serve as a reference for Ph.D. students and young researchers.

The book is organized in four parts: thermodynamics, kinetics, molecular structure and stability, and biophysical methods. To limit theoretical treatments to biologically relevant systems, we left out quantum mechanics and statistical thermodynamics. Each individual chapter is followed by a set of problems that should illustrate the principles treated in the chapter and put them into a biochemical context. The commented solutions to the problems are collected in a separate Solutions Manual. References for the work cited in the text are summarized at the end of each chapter, together with advanced literature that provides additional information and may serve as a starting point for delving into a particular subject.

The first part provides background on thermodynamic principles. Early on in the text, we have tried to illustrate the relevance of these thermodynamic concepts for biochemical systems. Electrochemistry is also treated in this part because it is presented from a thermodynamic point of view.

The second part presents kinetics in a somewhat non-traditional way. We have chosen to start with pre-steady state kinetics. This approach allows us to introduce many basic principles, which we can then build upon in the subsequent treatment of the more traditional steady-state enzymology. We also included a rather complete derivation of integrated rate laws and velocity equations.

In the third part, molecular interactions relevant for biological molecules are reviewed in light of the thermodynamic principles from the first part of the book. The principles of protein and nucleic acid structures are presented, striving for a balance between comprehensiveness and a focus on aspects that are important for life sciences. This part assumes a basic knowledge of the concepts of chemical bonds, orbitals and hybridization.

The fourth part deals with biophysical techniques and their applications to the study of biological macromolecules and their interactions. The selection naturally reflects our own preferences and areas of expertise. We included widely used techniques that should be accessible to most researchers when trying to answer a particular biochemical question. Standard biochemical methods, such as gel electrophoresis and chromatography, have been omitted. For the techniques most frequently used in a biophysical laboratory, a chapter on "potential pitfalls" may help to avoid common mistakes. Many well-established methods, among them fluorescence microscopy and mass spectrometry, have been developed much further in recent years. We incorporated these novel developments and include specific application examples.

Cross-references within and between all parts of the book serve to emphasize common themes and to highlight recurrent principles. Boxes throughout the main text illustrate the relevance of a particular topic, provide additional information on applications, or show derivations of formulas, but may be skipped without breaking the flow of reading.

The appendix is specifically tailored to the requirements of the four main chapters. We only list constants and units discussed in the text, and restrict the treatment of mathematical concepts to operations that are used in the book. The appendix is meant to be a convenient reference, not a substitute for a thorough training in mathematics.

We hope that this book is an enjoyable read, provides inspiration, and motivates students and researchers alike to dive deeper into biophysical chemistry.

Dagmar Klostermeier
Markus G. Rudolph
Muenster, June 2016

Preface to the second edition

More than five years have passed since the first edition of *Biophysical Chemistry* was published. The success of the book was unexpected for us, and we thank everyone who provided positive feedback. We are particularly grateful to those readers who spotted mistakes, and who commented on how the book could be made an even better teaching resource and reference.

Although the basic principles of biophysical chemistry have of course not changed, there have been scientific advances and technical developments in various fields since the first edition. We therefore felt that it was time for an update.

The overall structure of the book, with its four main Parts on Thermodynamics, Kinetics, Molecular structure and stability, and Methods, remains unaltered, but we have strengthened the connection between these Parts. In all four Parts, Chapters have been re-structured. In Part I, the very short Chapter 1 on Systems and their surroundings was combined with Chapter 2 on State functions and the laws of thermodynamics. In Part IV, Chapter 23 on Imaging and microscopy was split into Fluorescence imaging and microscopy (Chapter 25) and Electron microscopy (Chapter 26). The Section on cryo-electron microscopy was expanded to reflect its increasing importance in structural biology. The Section on Scanning probe microscopy was combined with the Chapter on Force measurements into a new Chapter 27 on Scanning probe microscopy and force measurements. The Sections on sequence analysis and molecular modeling in the Section on Computational biology were split. Protein and nucleic acid sequence analysis is now integrated into the corresponding Chapters 18 and 19 on these macromolecules. The Section on Molecular modeling formed the basis for the revised and expanded Chapter 20 on Macromolecular modeling.

New content has been included. In Part I, we added Chapter 3 on Statistical thermodynamics and its relevance for biological systems. A box on non-equilibrium thermodynamics was also added. In Part II, we now provide a step-by-step derivation for all integrated rate laws. New Chapters on the derivation of integrated rate laws by matrix algebra (Chapter 9) and on Single-molecule kinetics (Chapter 12) are included. Chapter 13 on enzyme kinetics now contains new Sections on IC_{50} values and their relation to K_d values, and on slow-onset inhibition. Chapter 16 on Principles of catalysis now treats hetero- and homogeneous catalysis, and emphasizes the common principles. Part III was revised to put more emphasis on the biophysical principles underlying structure formation and stability. In Chapter 18, the hydrophobic effect was moved from a Box to the main text to reflect its importance in protein stability. In Part IV, new Boxes introduce newly developed techniques that have found their way into standard applications, such as microscale thermophoresis in Chapter 21, mass photometry in Chapter 23, spinning disc-, MinFlux-, and expansion microscopy in Chapter 25. Chapter 20 now includes the basics of machine learning and the use of AlphaFold in biophysics. Serial crystallography and X-ray free-electron lasers, previously treated in a Box, now form a new Section in Chapter 24. Throughout the book, text and Figures were updated, inconsistencies in style were removed, and new Figures were added. The literature cited was updated.

Finally, the Mathematical Appendix was updated on statistics, implicit derivatives, and matrix algebra. Cross-references were added to the Sections where the topic is put into practice. The Tables on prefixes, units, and constants, previously the final Chapter, are now printed on the inside covers for quick reference.

Last, but not least, we added a number of end-of-chapter Questions, and updated the Solutions Manual.

We are confident that the book has improved with the changes we made, and we hope it continues to motivate and inspire current and future generations of students and scientists.

Acknowledgments to the first edition

We are grateful to our colleagues and friends who have taken the time to carefully read large parts of the drafted manuscript: David Banner, Hans-Joachim Galla, Walter Huber, Jürgen Köhler, John Ladbury, Erwin Peterman, Jochen Reinstein, Paul Rösch, Franz-Xaver Schmid, Bernhard Schmidt, Joachim Seelig, Heinz-Jürgen Steinhoff, Dmitri Svergun, and Ian Wilson. Your constructive comments on how to improve the text, both scientifically and educationally, are much appreciated. A special "Thank You" goes to David Banner for proofreading of the entire manuscript. Of course, any remaining errors in the book remain the sole responsibility of the authors.

We also thank our editors at CRC Press, Francesca McGowan and Emily Wells, for their commitment and constructive support, and for swiftly answering any queries that came up during the writing of this book.

Acknowledgments to the second edition

Over the years, we received valuable feedback on the book from students, fellow teachers, and scientists, which has been incorporated into this second edition you now hold in your hands. We thank Josep A. Biosca, Thorben Cordes, Ute Curth, Óscar Herrera Chacon, Birgit Jaitner, John Jumper, Boris Musset, Andreas Reichert, and Gerard Marriott for pointing out mistakes and their feedback on how to improve the book, and Andreas Heuer and Lubomir Kovacik for proofreading individual Chapters. We also thank Danny Kielty from CRC Press for keeping an eye on our progress and Rebecca Hodges-Davies for her encouragement and her enthusiastic support.

Authors

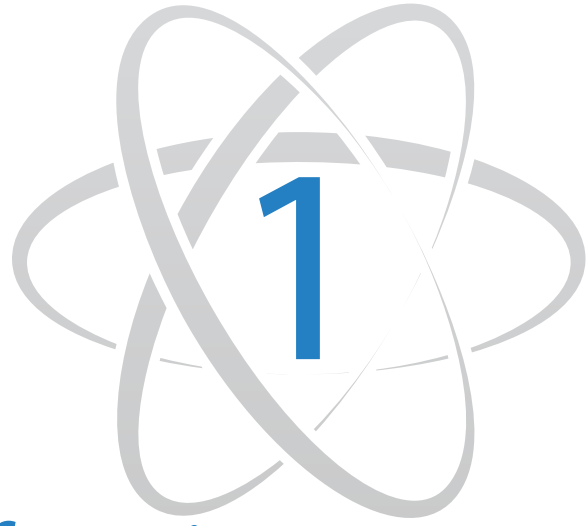
Dagmar Klostermeier studied Biochemistry and holds a PhD in Biochemistry. After her postdoctoral training with David Millar at The Scripps Research Institute in La Jolla, USA, she was a junior research group leader at the University of Bayreuth, Germany, and a Professor for Biophysical Chemistry at the University of Basel, Switzerland. In 2011, she became Professor for Biophysical Chemistry at the University of Muenster, Germany, and is teaching Physical and Biophysical Chemistry. Her research aims to decipher the mechanisms of ATP-driven molecular machines by biophysical techniques, including single-molecule fluorescence microscopy.

Markus G. Rudolph studied Biochemistry and holds a PhD in Biochemistry. He completed his postdoctoral training in Structural Biology in the laboratory of Ian A. Wilson at The Scripps Research Institute in La Jolla, USA, before leading his own research group at the University of Goettingen, Germany. In 2006, he moved to Hoffmann-La Roche Ltd. in Basel, Switzerland, as a researcher in Structural Biology and Biophysics. His main research interests are the structural biology and catalytic mechanisms of enzymes, and structure-based drug design.



PART I
Thermodynamics

The field of thermodynamics (gr. *thermos*: heat, *dynamis*: force) is centered around energy in all forms. It refers to the analysis of the available energy, the distribution of energy, and the conversion of energy from one form into the other. Thermodynamic considerations are invaluable for the understanding of the importance of energy in biological systems, its storage and availability, and the energetic costs involved in maintaining cellular functions and stability. The thermodynamic framework is based on four laws, simple rules that have been derived empirically from observing the real world. These laws and concepts allow a quantitative description of complex biological systems, and enable us to understand and, more importantly, to predict their behavior in specific circumstances.



1 Systems, state functions, and the laws of thermodynamics

1.1 SYSTEMS AND THEIR SURROUNDINGS

Before we begin to construct the building of thermodynamics, we will have to start with a set of definitions. A *system* is the part of the universe we are currently focusing on, and that we are attempting to describe quantitatively. Systems are separated from the rest of the universe by walls that allow them to interact with their *surroundings*. By surroundings we refer to that part of the rest of the universe that interacts with the system (Figure 1.1). An *isolated system* is separated from its surroundings by walls that do not allow exchange of heat or matter. A *closed system* exchanges heat with its surroundings, but does not exchange matter, and an *open system* can exchange both heat and matter with its surroundings. Walls that allow heat exchange are called *diathermic*, whereas walls that do not allow the exchange of heat are called *adiabatic*.

These definitions imply that all living organisms are open systems: they exchange heat and matter with their surroundings. Open systems are not so easy to describe quantitatively, however. We will therefore start with isolated and closed systems in order to illustrate the underlying concepts, before we move to open systems (Section 1.7).

Energy exchanged between a system and its surroundings is always counted from the perspective of the system. A negative value means energy flows from the system into the surroundings: the energy of the system is reduced. A positive value means that energy flows into the system, and the energy of the system is increased (Section 1.3.2).

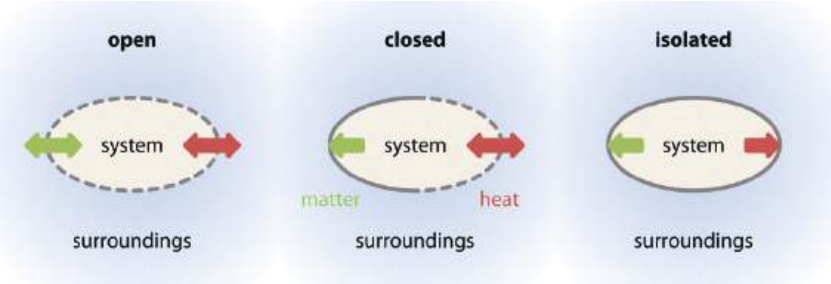


FIGURE 1.1: Systems and surroundings: open, closed, and isolated systems. A system is the part of the universe we want to understand (pale yellow). It is in contact with its surroundings (blue). System and surroundings are separated by walls. Open systems exchange heat (red) and matter (green) with their surroundings, closed systems exchange heat, but not matter, and isolated systems exchange neither heat nor matter with their surroundings. Open and closed systems are separated from their surroundings by diathermic (heat-permeable) walls; isolated systems are surrounded by adiabatic (thermally insulated) walls.

Now that we have defined different types of systems, we can start to dive into the laws of thermodynamics. First, we have to familiarize ourselves with the properties of state functions. *State functions*, as their name suggests, describe the state of a system as a function of individual *state variables* (sometimes referred to as state parameters, or also as state function). We will see in the following that state functions are important and very convenient tools to describe changes in the state of a system.

1.2 STATE VARIABLES AND STATE FUNCTIONS

The physical state of matter is described by a number of parameters, termed *state variables*. For example, the state of a gas is characterized by its volume, temperature, and pressure. The gas volume is delimited by the volume of the container in which the gas is kept. The temperature defines the thermal energy of the gas, and the velocity with which the gas particles move randomly within the container. The pressure is caused by the moving gas particles hitting the container walls. The *ideal gas law* describes the connection between the state variables volume V, temperature T, and pressure p for an ideal gas and connects them to the amount of substance n (eq. 1.1):

$$pV = nRT$$

or eq. 1.1

$$p = \frac{nRT}{V}$$

with the pressure p in Pa, the volume V in m^3, the amount of substance n in mol, and the absolute temperature T in K. R, the general gas constant, is the proportionality constant between the products pV and nT, and has a value of R = 8.3143 J mol^{-1} K^{-1}. Note that throughout this book, we use the standard international units (SI units). The ideal gas law defines the state of an ideal gas unambiguously: if three out of the four state variables p, V, T, and n are known, the fourth is determined automatically by eq. 1.1. Similarly, any thermodynamic *state function* defines the state of a system unambiguously, without leaving room for interpretation. If n–1 of its n parameters are known, the remaining value can be directly calculated from the state function. Note that the units of the expressions on the left-hand and right-hand side of eq. 1.1 are units of energy (J). We will come back to the energy described by pV and nRT in Section 1.3.4.4.

State functions derive their name from the property that they only depend on the current state of a system. As an important consequence of this property, state functions are independent of the pathway by which this state has been reached. This formal definition has very intriguing implications,

and we will see later on that this property of state functions simplifies the description of systems tremendously. State functions are called *intensive state functions* when they are independent of the system size (*e.g.* temperature T or pressure p). *Extensive state functions*, on the other hand, depend on the system size (*e.g.* mass m, volume V, internal energy U, enthalpy H, entropy S, free energy G). An extensive state function can be normalized to the size of the system, *e.g.* by dividing through the amount of substance, the volume, or the mass. The resulting number is independent of the system size, and thus falls under the definition of an intensive state function. Normalized state variables are *specific* (normalized by mass m), *molar* (normalized by amount of substance n), or *density* (normalized by volume V). Examples are the specific heat capacity C, the molar enthalpy H_m, molar entropy S_m, molar free energy G_m and chemical potential μ, or the density ρ.

We have noted before that state functions are independent of the pathway by which a certain state is reached. The highly convenient consequence is that changes in state functions during certain processes can be expressed simply by the differences between their final value and their initial value (Figure 1.2). This concept can be illustrated using the state variable temperature T as an example. When energy is transferred to a system, such that the starting temperature $T_{initial}$ of the system increases to the final value T_{final}, the change in temperature, ΔT, can be calculated as the difference between final and initial values of T (eq. 1.2):

$$\Delta T = T_{final} - T_{inital} \qquad \text{eq. 1.2}$$

Here, ΔT refers to a large, measurable change of the temperature. In general, we refer to large, measurable changes in a state function X by ΔX. For state functions, this change can always be calculated from the final and initial values (eq. 1.3):

$$\Delta X = X_{final} - X_{inital} \qquad \text{eq. 1.3}$$

With dX, on the other hand, we refer to a small, infinitesimal change in the state function X. In contrast, with δY we refer to a small, infinitesimal change in a function Y that is NOT a state function.

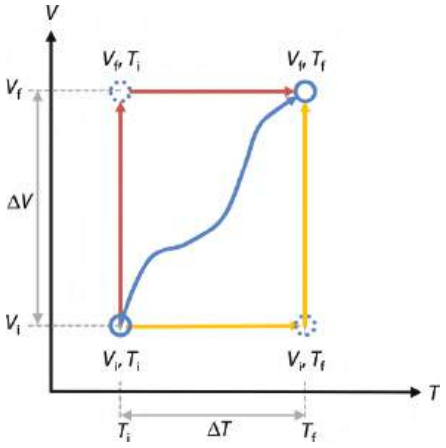


FIGURE 1.2: Properties of state functions. A system changes from an initial state with temperature T_i and a volume V_i to a final state with temperature T_f and volume V_f. Such a change in state can follow different paths: along the red path, a change from initial to final volume at constant (initial) temperature is followed by a change from the initial to the final temperature at constant (final) volume. Along the yellow path, the temperature changes first from its initial to the final value at constant (initial) volume, and the volume changes from its initial to its final value at constant (final) temperature. The blue path shows a pathway where temperature and volume change continuously throughout the transition. Temperature and volume are state functions, and their change in the transition can be calculated from the final and initial values: $\Delta V = V_f - V_i$ and $\Delta T = T_f - T_i$. ΔV and ΔT are therefore identical for all three pathways, and are generally independent of the pathway.

A system is in *equilibrium* when its state and the parameters that describe this state do not further change with time. A system is in *thermal equilibrium* if its temperature remains constant with time (Figure 1.3). It is important to note that this does not necessarily imply that there is no heat exchange; it just indicates that the heat entering and exiting the system is balanced, and the net temperature change is zero. We will revisit this principle of *dynamic equilibria* when we discuss the kinetics of reversible reactions (Sections 8.1, 13.6.1).

Two systems are in thermal equilibrium when they have the same temperature (that is constant over time). If two systems, A and B, are in thermal equilibrium, and at the same time, the system B is in thermal equilibrium with system C, we can conclude that systems A and C must also be in thermal equilibrium. This statement is the 0^{th} law of thermodynamics: if out of three systems A, B, and C, systems A and B, and systems B and C are in pairwise equilibrium, then systems A and C are also in thermal equilibrium (Figure 1.3). This is a recurrent principle that also holds for chemical equilibria, and has important implications for coupled chemical reactions in biology (Section 2.4).

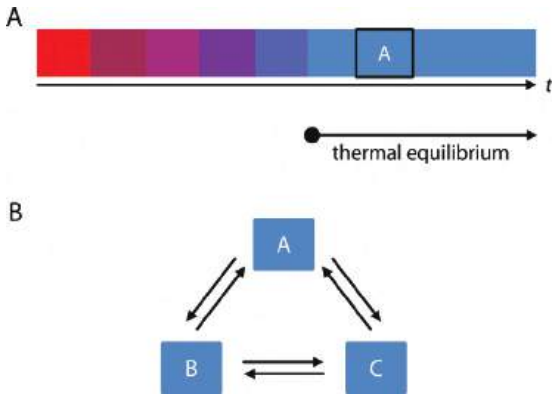


FIGURE 1.3: Thermal equilibrium and the 0^{th} law of thermodynamics. A: A system is in thermal equilibrium when its temperature remains constant over time. If we let a hot system (red) equilibrate with its cooler surroundings, the system will eventually reach a constant temperature (blue) and thermal equilibrium (arrow). B: When system A is in thermal equilibrium with systems B and C, then systems B and C are also in thermal equilibrium. This relationship is the 0^{th} law of thermodynamics.

1.3 THE INTERNAL ENERGY *U* AND THE FIRST LAW OF THERMODYNAMICS

1.3.1 Internal energy, heat, and work

Any system contains a certain amount of energy, which is called its *internal energy*. The internal energy U is the sum of all forms of energy that are stored in the system. It can be regarded as the sum of the kinetic energy of all particles in a system plus the sum of their potential energy. Internal energy consists of two major components, *work w* and *heat q*. Heat is the form of energy associated with random movement of particles. In contrast, work is energy associated with ordered movement. Work can be subdivided into mechanical work, electrochemical work, expansion work, and other forms (Figure 1.4). In the following, we will introduce the laws of thermodynamics and define state functions by focusing on expansion work performed during state changes of ideal gases.

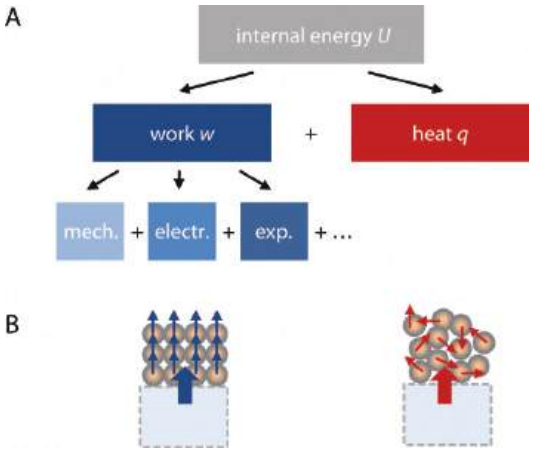


FIGURE 1.4: Internal energy, heat, work, and different types of work. A: The internal energy of a system is divided into heat *q* and work *w*. Work can be further subdivided into mechanical (mech.), electrical (electr.), expansion (exp.) work, and other forms (…). B: Heat is associated with random movement of particles, work is associated with ordered movement.

1.3.2 The first law of thermodynamics

Changes in the internal energy of a system, dU, can be brought about by adding or withdrawing heat (δq), or by performing work on the system or having the system perform work (δw) on the surroundings:

$$dU = \delta w + \delta q \qquad \text{eq. 1.4}$$

The internal energy of a system increases when energy is added, either by performing work on the system or by adding heat (or both). The internal energy of a system decreases if the system performs work and/or transfers heat to its surroundings. The notation in eq. 1.4, using dU, but δw and δq, already implies that, in contrast to the state function U, w, and q are not state functions. Indeed, despite U being a state function, the distribution of changes in internal energy to work and heat depends on the pathway of the respective process, and hence w and q are not state functions. We will see later that work and heat, both forms of energy, are not equivalent. Work can be converted into heat, but heat cannot be completely converted into work (Section 1.5.2).

Changes in work and heat are always formulated from the perspective of the system (Figure 1.5): $\delta w > 0$ means work is performed on the system, leading to an increase in its internal energy. $\delta w < 0$ means that the system performs work on the surroundings, leading to a decrease in internal energy of the system. Correspondingly, a change in heat $\delta q > 0$ means that energy in the form of heat is added to a system, whereas $\delta q < 0$ means that heat is removed from the system, leading to an increase or decrease of the internal energy, respectively (eq. 1.4).

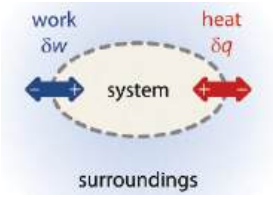


FIGURE 1.5: System-centered sign convention of heat and work. Accounting of energy is always done from the perspective of the system. Heat transferred to the system enters the balance with a positive sign ($+\delta q$), heat transferred from the system to its surroundings is accounted for with a negative sign ($-\delta q$). Work performed on the system has a positive sign ($+\delta w$), and work performed by the system on its surroundings has a negative sign ($-\delta w$).

1.3.3 The ideal gas: a convenient system to understand thermodynamic principles

We have introduced before the ideal gas law which describes the relation between the state variables pressure, volume, temperature, and amount of substance (eq. 1.1). The ideal gas law tells us that the product of pressure p and volume V increases with the temperature T of the gas (Figure 1.6). At a constant temperature (*isothermal* conditions), the product pV will be constant. This relationship was found experimentally, and is called the *Boyle-Mariotte law*. Curves that describe p as a function of V for a constant temperature T are called *isotherms* (Figure 1.6). If we keep the pressure constant (*isobaric* conditions), then the volume V is proportional to the temperature T, a relationship known as *Charles' law* or the *first law of Gay-Lussac*. Plotting V as a function of T for different pressures gives a set of lines that are termed *isobars*. At constant volume (*isochoric* conditions), the pressure p will increase proportionally with the temperature T. This relationship is called the *second law of Gay-Lussac*. The curves that describe p as a function of T for a constant volume are called *isochors*. Finally, the proportional relationship of the volume V to the amount of substance n at constant temperature and pressure (isothermal and isobaric conditions) is *Avogadro's law*. The ideal gas law combines these individual pairwise relationships between temperature, pressure, and volume, and the amount of substance, and introduces the proportionality constant between pV and nT, the general gas constant R.

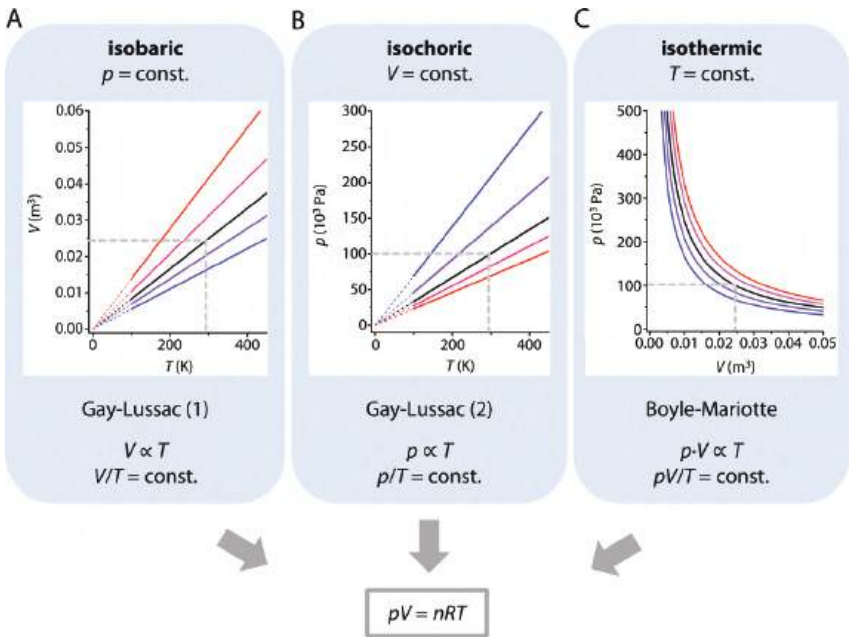


FIGURE 1.6: The ideal gas law and its partial laws. A: The first law of Gay-Lussac states the proportionality of the volume V and the temperature T of an ideal gas under isobaric conditions (p = const., increasing pressure from red through black to blue). The broken lines indicate the extrapolation to the absolute temperature minimum. B: The second law of Gay-Lussac states the proportionality of the volume V and the temperature T of an ideal gas under isochoric conditions (V = const.; decreasing volume from red through black to blue). The broken lines indicate the extrapolation to the absolute temperature minimum. C: The law of Boyle and Mariotte states the proportionality of the product of the pressure p and the volume V and the temperature T of an ideal gas under isothermal conditions (T = const.). All three individual laws are integrated into the ideal gas equation. The ideal gas law also contains Avogadro's law: the same number of particles (*i.e.* amount of substance n) of a gas takes up equal volumes at constant pressure p and temperature T.

The proportionality of V and T has an important consequence because it defines a lower limit for the temperature scale. The volume V reaches zero at a certain temperature, where all isobars intersect (Figure 1.6). The volume is always positive ($V > 0$), and lower temperatures are therefore not possible. This temperature is thus the absolute zero of the thermodynamic temperature scale ($T = 0$ K). Because of the finite volume of atoms and molecules, a volume of zero cannot be reached, and the absolute minimum in temperature also cannot be reached. We will revisit this statement in the context of the third law of thermodynamics (Section 1.5.4).

The ideal gas law, and particularly the Boyle-Mariotte law, has important implications for scuba divers (Box 1.1). Many fish exploit the properties of gases and use gas reservoirs to control their buoyancy (Box 1.2).

BOX 1.1: UNDER PRESSURE: IMPLICATIONS OF THE IDEAL GAS LAW FOR DIVING.

The mean air pressure at sea level fluctuates around $1.013 \cdot 10^5$ Pa (1.013 bar). The air pressure at sea level is caused by the air layers above. Water has a higher density than air, and under water we therefore experience a higher pressure than above the surface. At a depth of 10 m, the water column above us leads to an increase in pressure by another $1.013 \cdot 10^5$ Pa, and every additional ten meters will add approximately $1.013 \cdot 10^5$ Pa. Thus, we experience twice the normal pressure at 10 m (once from the air above us, once from the water), three times the normal pressure at 20 m, four times the normal pressure at 30 m, and so on. For physiological processes, the increased pressure is not a problem: most of our body is water, and water is almost incompressible. The most critical organ is our lung, a large air-filled reservoir. Following the Boyle-Mariotte law ($pV = \text{const.}$), the air in our lungs is compressed while we descend during our dive and the external pressure increases, and expands upon ascent with decreasing pressure. As long as we take a deep breath to fill our lungs at the surface, then dive while holding our breath, and ascend again, the compression of air in our lungs is not problematic at all: we start our dive with full lungs, the air becomes compressed during the dive, and expands to its original volume when we reach the surface. The problems start when we extend our under-water time by breathing compressed air. Scuba divers carry a supply of compressed air to be able to breath under water. The air in their tanks is pressurized to typically $2 \cdot 10^7$ Pa (200 bar), and is down-regulated to environmental pressure in two stages for breathing. With every breath under water, divers fill their lungs with air at the pressure experienced at that particular depth. Breathing 20 meters below the surface means breathing air of three times the pressure (and a third of the volume) it would have at the surface. When we ascend to finish our dive, the surrounding pressure decreases with every meter of ascent, and the air in our lungs expands. If we have filled our lungs at 20 m and ascend without breathing, the volume will increase three times, which is far beyond the capacity of our lungs. The lung tissue is not able to accommodate this volume, and will eventually rupture, which leads to a life-threatening pulmonary barotrauma. To counteract the expansion of air in our lungs, we have to ascend slowly to ensure that the air expands slowly, and we have to breathe constantly. Every exhalation removes high-pressure air from the lungs, and every inhalation fills the lungs with air of lower pressure. The pressure of the air in the lungs and the external pressure are thereby equilibrated during the ascent, avoiding pressure differences, air expansion, and lung rupture. Failure to ascend slowly and to constantly release air from the lungs during the ascent can cause fatal lung injuries, and the first lesson scuba divers learn is: never hold your breath – a life-saving lesson from the Boyle-Mariotte law.

BOX 1.2: WHY FISH DO NOT SINK: THE IDEAL GAS LAW AND BUOYANCY.

Many fish species contain a swim bladder, a gas-filled compartment. These fish can adjust the amount of gas in the swim bladder to control their buoyancy (Figure 1.7). Adding gas to the swim bladder from the bloodstream (or the gut) causes the fish to ascend, removing gas causes the fish to sink. When a fish with neutral buoyancy swims upward, the gas in the swim bladder expands because of the decreasing pressure. To regain neutral buoyancy, gas needs to be absorbed from the bladder. When the fish swims to deeper waters, the gas in the swim bladder is compressed by the increasing pressure, and gas needs to be added to prevent the fish from sinking further. Fish with a swim bladder can remain at constant depth without effort. In contrast, fish without a swim bladder, such as sharks, need to constantly swim to prevent sinking.

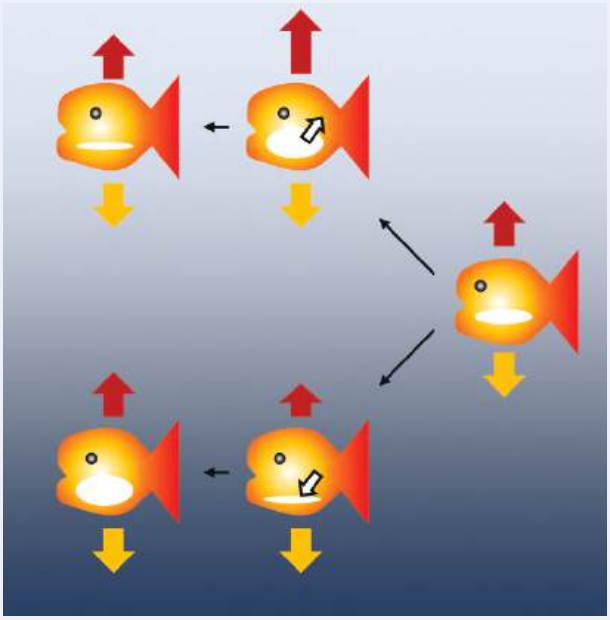


FIGURE 1.7: Buoyancy control with a swim bladder. The swim bladder (white) is a gas-filled reservoir that allows fish to control their buoyancy. When a fish with neutral buoyancy ascends (top), the air in the swim bladder expands, leading to positive buoyancy. To prevent further ascent and regain neutral buoyancy, the fish needs to decrease the amount of air in the bladder (white arrow). When a fish with neutral buoyancy descends (bottom), the air in the swim bladder is compressed, leading to negative buoyancy. To prevent further descent and regain neutral buoyancy, the fish increases the amount of air in the bladder (white arrow).

The same principle is used by divers who wear an air-filled vest as a buoyancy control device. The device is connected to the compressed air supply of the diver. Increasing the amount of air in the vest causes a diver with neutral buoyancy to ascend or prevents a diver with negative buoyancy from further sinking upon compression of the air at increasing depth. Conversely, releasing air from the vest through a valve makes a diver with neutral buoyancy descend, and prevents an ascending diver from uncontrolled ascent due to expansion of the air with decreasing pressure.

1.3.4 Changes in the state of an ideal gas

We will now look at state changes of an ideal gas under different conditions, and calculate the changes in work and heat that are associated with these processes.

1.3.4.1 Irreversible isothermal expansion and compression

First, we start with an ideal gas in an initial state characterized by pressure and volume p_1, V_1 and a certain temperature T (Figure 1.8). The gas is in a cylinder, closed off with a moveable piston. From the starting conditions (p_1, V_1, T), we now perform an *irreversible expansion* to a state with p_2, V_2, T: we instantaneously move the piston upward, such that the volume the gas can fill is increased. At the external pressure p_2, the gas will then expand from the initial volume V_1 into the volume V_2. We perform this expansion at constant temperature T by coupling the gas to a large heat reservoir. According to the Boyle-Mariotte law, the product pV for the initial state (p_1, V_1) and the final state (p_2, V_2) is constant under these isothermal conditions. From the increase in volume, $V_2 > V_1$, we can therefore conclude that the pressure must decrease during the change in state, with $p_2 < p_1$.

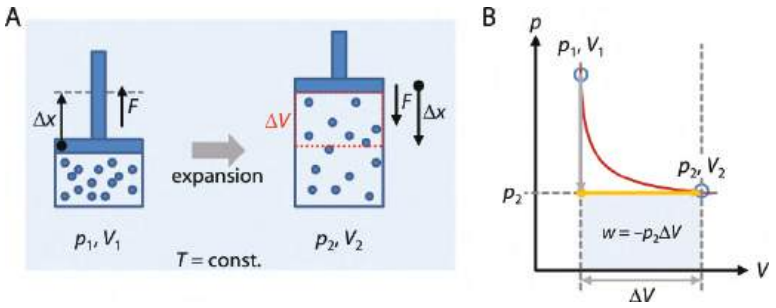


FIGURE 1.8: Irreversible expansion. A: An ideal gas is kept in a container with a moveable piston at pressure p_1 and volume V_1. The temperature T is maintained constant by coupling the gas to a large heat reservoir (blue). We now rapidly move the piston by Δx, providing a larger volume for the ideal gas to occupy. The ideal gas will expand into the volume ($V_2 > V_1$). The product of p and V is constant under isothermal conditions, hence $p_2 < p_1$. B: The states p_1, V_1 and p_2, V_2 are on the isotherm (red line). The irreversible expansion (orange arrow) does not follow the isotherm, however: the pressure is reduced from p_1 to p_2 instantaneously (gray arrow), and the gas experiences the external pressure p_2 throughout the expansion. The work performed during the expansion is $w = -p_2\Delta V$ (blue area, see text).

We can calculate the work associated with this expansion from the mechanical work w_{mech} that is performed to move the piston to its new position. This work is the product of the force F with which the piston is pushed out and the distance Δx by which it moves:

$$w_{mech} = F \cdot \Delta x \qquad \text{eq. 1.5}$$

The force can be expressed as the product of the pressure p on the piston and its area A. We have said above that we move the piston instantaneously. As a consequence, the complete change in state is performed under the final pressure p_2, and this is the pressure that acts on the piston throughout the expansion of the gas (Figure 1.8). We obtain for the work w_{mech}:

$$w_{mech} = p_2 \cdot A \cdot \Delta x \qquad \text{eq. 1.6}$$

By summarizing the product of the area A and Δx as the change in volume of the gas, ΔV, the work for moving the piston becomes

$$w_{mech} = p_2 \cdot \Delta V \qquad \text{eq. 1.7}$$

The work w associated with the irreversible expansion of the gas is of the same amount, but has the opposite sign:

$$w = -p_2 \Delta V \qquad \text{eq. 1.8}$$

The negative sign of the work w reflects the system-centered perspective: the system performs work on the surroundings by moving the piston upward, therefore, the work w has to be negative. Conversely, we can calculate the work associated with an irreversible compression from p_2, V_2 to p_1, V_1 from the product of the final pressure p_1, which is experienced throughout the compression, and the change in volume $\Delta V = V_1 - V_2$. To compress a gas, the piston needs to be pushed down, and we need to perform work on the system. The compression work therefore has a positive sign, and is stored in the system. The positive sign is directly obtained from eq. 1.8 because the volume change ΔV for the compression is negative, making w positive.

During an irreversible expansion, the gas performs work on the surroundings, which would lead to a decrease in the internal energy of the system. However, if we perform the expansion of the gas under isothermal conditions, maintaining a constant temperature by coupling the system to a (large) heat reservoir such as a giant water bath, heat will be taken up by the gas upon expansion to compensate for the loss in internal energy due to the work performed on the surroundings. Conversely, heat will be released (*i.e.* taken up by the water bath) upon compression to compensate for the increase in internal energy because of the work performed on the system by the surroundings. The transferred heat q during an irreversible isothermal expansion or compression is thus

$$q = -w = p_2 \Delta V \qquad \text{eq. 1.9}$$

Overall, the change in internal energy of the gas is then

$$\mathrm{d}U = q + w = p_2 \Delta V - p_2 \Delta V = 0 \qquad \text{eq. 1.10}$$

Thus, during an isothermal, irreversible expansion, the internal energy of the gas does not change.

1.3.4.2 Reversible isothermal expansion and compression

We now consider the case that instead of instantaneously moving the piston to the final position in one go, we perform the same isothermal expansion under *reversible* conditions by slowly moving the piston by infinitesimally small increments $\mathrm{d}x$. The volume then also changes in very small, infinitesimal steps $\mathrm{d}V = A \cdot \mathrm{d}x$. Under these conditions, the gas pressure can equilibrate to the pressure according to the ideal gas law at each point of the expansion (Figure 1.9). We can now calculate the mechanical work on the piston by integrating the associated force with the change in position, $\mathrm{d}x$.

$$w_{\mathrm{mech}} = \int F \, \mathrm{d}x \qquad \text{eq. 1.11}$$

By analogy to the considerations for the irreversible expansion, the expansion work performed by the gas during the reversible expansion is now also described by an integral

$$w = \int_{V_1}^{V_2} -p \, \mathrm{d}V \qquad \text{eq. 1.12}$$

Here, p is the current pressure, which we can substitute for nRT/V from the ideal gas law (eq. 1.1). The integral can then be evaluated as

$$w = \int_{V_1}^{V_2} -p \, \mathrm{d}V = -nRT \int_{V_1}^{V_2} \frac{1}{V} \mathrm{d}V = -nRT \ln \frac{V_2}{V_1} \qquad \text{eq. 1.13}$$

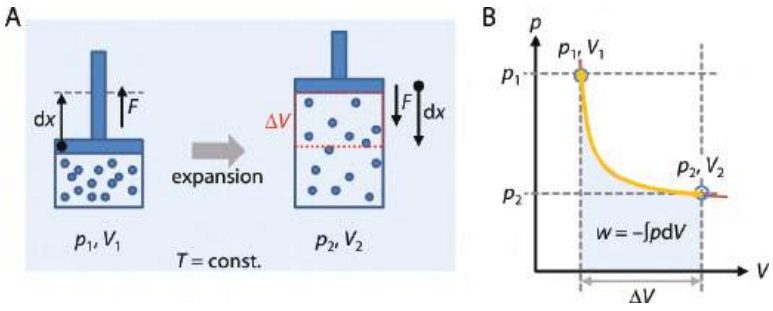


FIGURE 1.9: Isothermal expansion under reversible conditions. To expand an ideal gas reversibly, we move the piston up slowly in small, infinitesimal steps d*x*. At each point of the expansion process, the pressure *p* of the gas will equilibrate under the new conditions, and the volume will adjust according to the ideal gas law (orange arrow). We can calculate the work from the integral of –*p*d*V*, which corresponds to the blue area under the isotherm (red curve) between initial and final conditions. Note that the work performed by a reversibly expanding gas is larger than the work performed by an irreversibly expanding gas (Figure 1.8).

According to eq. 1.13, the work associated with a reversible isothermal expansion ($V_2 > V_1$, $\ln(V_2/V_1) > 0$) is negative; the system performs work on the surroundings. However, for a reversible isothermal compression from p_2, V_2 to p_1, V_1 ($V_1 < V_2$, $\ln(V_1/V_2) < 0$), the work according to eq. 1.13 is positive; work is performed on the system.

The work performed by the system on the surroundings during the reversible expansion would lead to a decrease in internal energy of the system. Again, under isothermal conditions heat is exchanged that is of the same magnitude as the work performed, but of opposite sign:

$$q = -w = nRT \ln \frac{V_2}{V_1} \qquad \text{eq. 1.14}$$

and the internal energy of the gas will remain constant during the expansion:

$$dU = q + w = nRT \ln \frac{V_2}{V_1} - nRT \ln \frac{V_2}{V_1} = 0 \qquad \text{eq. 1.15}$$

1.3.4.3 Comparison of reversible and irreversible changes of state

If we now compare the work associated with the irreversible (eq. 1.8) and reversible (eq. 1.13) expansions, we see immediately that the work performed by the system during an irreversible expansion is smaller than the work performed during a reversible expansion (Figure 1.9). This is caused by the constant pressure *p* during an irreversible expansion where the complete expansion occurs under the lower final pressure p_2. An irreversible compression, on the other hand, requires more work than the same compression under reversible conditions because the entire process occurs at the final higher pressure p_1 (Figure 1.10).

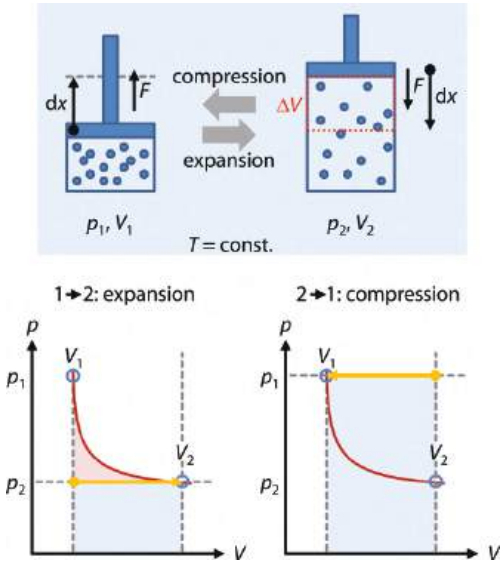


FIGURE 1.10: Work for irreversible and reversible expansion and compression. A gas expanding under irreversible conditions performs less work (blue area) than a gas that is expanding under reversible conditions (blue and light red area). Compressing a gas under irreversible conditions requires more work (blue area) than under reversible conditions (area under the red curve). The path for irreversible expansion and compression is indicated by the orange arrows. Reversible expansion and compression follow the isotherm (red curve).

Instead of expanding or compressing an ideal gas in one step from its initial to its final volume, we can perform the expansion or compression in several irreversible steps (Figure 1.11). If the number of steps is sufficiently large, the overall work for n irreversible steps will become equal to the work under reversible conditions. The reason for this is the smaller and smaller difference between the final pressure for each irreversible step and the pressure p under reversible conditions.

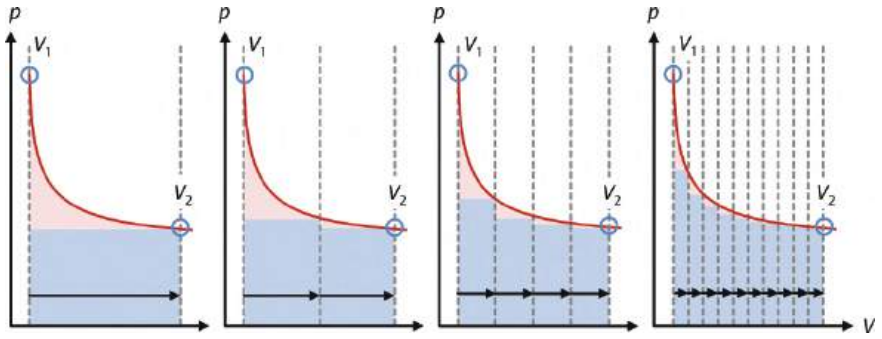


FIGURE 1.11: From irreversible to reversible expansion of an ideal gas. The red curve represents the p-V relationship according to the ideal gas law. We can expand an ideal gas from an initial volume V_1 to a final volume V_2 under irreversible or reversible conditions. The work performed by a reversibly expanding gas (area under the curve) is larger than the work performed by an irreversibly expanding gas (blue area). The excess of work performed by the reversibly expanding gas is highlighted in light red. We can now perform the same change of state from V_1 to V_2 in several irreversible steps (*e.g.* 2, 4, or 10 steps). For each individual irreversible expansion step, the work performed by the gas is indicated by a blue rectangle. The excess work performed by a reversibly expanding gas is indicated in light red. With more and more irreversible sub-steps, the difference between work performed under reversible and irreversible conditions becomes smaller and smaller. For an infinite number of steps, each step will be an expansion by an infinitesimal volume increment dV, which is the limiting case of an expansion under reversible conditions. The total work performed in an infinite number of irreversible expansion steps (sum of blue areas) will be identical to the work performed under reversible conditions (area under the curve).

From Figure 1.11, we see that the work performed during a reversible expansion constitutes an upper limit for the work performed by the system:

$$|w_{rev}| \geq |w_{irrev}| \qquad \text{eq. 1.16}$$

For the compression, on the other hand, the work performed under reversible conditions is a lower boundary for the work performed on the system:

$$|w_{rev}| \leq |w_{irrev}| \qquad \text{eq. 1.17}$$

In other words, for a real process that is not entirely reversible, the work w_{irrev} the system performs is smaller than w_{rev}, whereas the work we have to perform on a system is larger. This relation does not only apply to state changes of gases but to all processes: in general, we can have a system perform more work on the way from state 1 to state 2 if changes are reversible. We will come back to this in Section 1.5.2.

1.3.4.4 Adiabatic expansion and compression

So far, we have considered volume changes of an ideal gas under isothermal conditions, where the temperature of the gas remains constant because the gas is coupled to a large heat reservoir, such that heat released by the gas is taken up by the reservoir, and heat taken up by the gas is removed from the reservoir. We can also change the state of an ideal gas without allowing heat exchange with the surroundings. A change in state without heat exchange is called an *adiabatic process*. In this case, the temperature of the system will change. Upon expansion, the system cannot take up heat from the surroundings, and the temperature will decrease. Conversely, upon compression, no heat can be released, and the temperature of the system will increase. Adiabatic compression of air masses crossing a mountain range is the reason for warm winds in the valleys (Box 1.3).

BOX 1.3: ADIABATIC COMPRESSION OF AIR MASSES GENERATES DRY, WARM WINDS.

When winds move air toward high mountains, the air masses ascend to higher levels (Figure 1.12). They thereby reach regions with lower air pressure, and expand. The air does not exchange heat with its surroundings during ascent, and the expansion process is nearly adiabatic, resulting in cooling of the air. The cooler air cannot hold as much humidity as the warm air, which leads to cloud formation and relief rainfall. The now dry air rises further and crosses the highest ridge. On the other side of the mountains, the air descends rapidly, and again this process is adiabatic. The increase in pressure leads to compression of the air masses, and to an increase in temperature. This warmer air can take up more humidity, and will dry the area along its path.

(*Continued*)

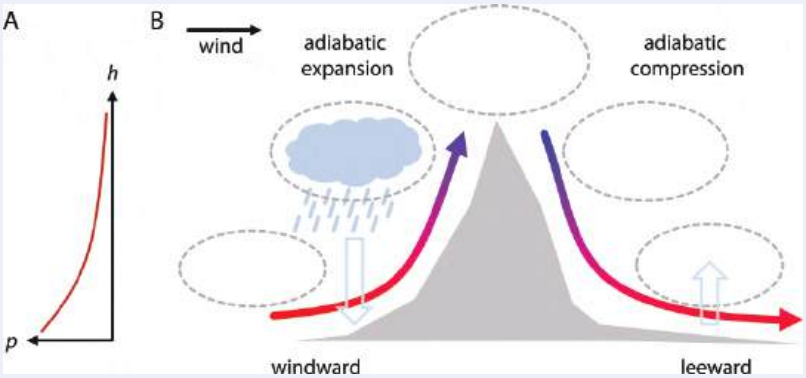


FIGURE 1.12: Adiabatic expansion and compression and dry fall winds. A: The air pressure p decreases exponentially with increasing height h above sea level. B: Air masses (gray) are blown toward a mountain range by winds, and ascend on the windward side. During adiabatic expansion, the air becomes cooler (arrow) and can hold less humidity, leading to cloud formation and rainfall. On the wind-distant (leeward) side, the dry air is compressed adiabatically during descent. The air warms and takes up humidity from the ground (blue arrow).

Dry, warm fall winds frequently occur on the northern side of the Alps (Föhn), and on the wind-distant side of other mountain ranges, such as the Sierra Nevada (Santa Ana winds), the Rocky Mountains (Chinook), or the Andes (Chanduy, Puelche, Zonda). Because of their dryness, these winds cause health and environmental problems, and promote wildfires and desert formation.

To quantify the work performed during an adiabatic expansion, we have to take into account that now two state variables are changing: the temperature T and the volume V. In this case, the property of a state function comes in handy: the internal energy of the system only depends on the initial and final states, not on the pathway on which the system changes from one state to the other. We can therefore describe the adiabatic expansion by splitting it into two steps: an isothermal expansion from V_1 to V_2 at constant temperature T_1, followed by an isochoric change in temperature from T_1 to T_2 at a constant volume V_2 (Figure 1.13).

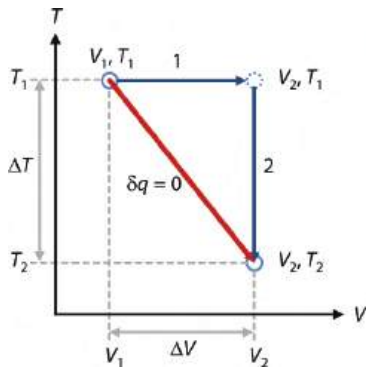


FIGURE 1.13: An adiabatic expansion is an expansion at constant temperature, followed by a change in temperature at constant volume. In an adiabatic expansion, the volume of the gas increases from V_1 to V_2. No heat is exchanged, and the temperature of the gas will therefore decrease from T_1 to T_2 during the expansion. To quantitatively describe the work associated with this adiabatic change in state, we can make use of the fact that changes in state functions are independent of the pathway. This independence of the pathway allows us to dissect the adiabatic expansion (red arrow) into two steps (blue arrows): an isothermal expansion at a constant temperature T_1 (1), followed by an isochoric decrease in temperature from T_1 to T_2 at a constant volume V_2 (2).

The change in internal energy dU during an adiabatic expansion is thus the sum of the changes in internal energy in the two steps, the isothermal expansion, and the isochoric change in temperature:

$$\mathrm{d}U = \mathrm{d}U_{\mathrm{T}_1,\mathrm{V}_1\rightarrow\mathrm{T}_1,\mathrm{V}_2} + \mathrm{d}U_{\mathrm{T}_1,\mathrm{V}_2\rightarrow\mathrm{T}_2,\mathrm{V}_2} \qquad \text{eq. 1.18}$$

We can express the change in internal energy dU in the two steps as

$$\mathrm{d}U = \left(\frac{\partial U}{\partial V}\right)_{\mathrm{T}} \mathrm{d}V + \left(\frac{\partial U}{\partial T}\right)_{\mathrm{V}} \mathrm{d}T \qquad \text{eq. 1.19}$$

where the first term describes the change in U with the change in volume at constant temperature (isothermal expansion), and the second term corresponds to the temperature change at constant volume (isochoric process). The terms in parentheses are termed *partial derivatives* (see Section 31.6.2), because the differentiation is only performed with respect to one variable. These partial derivatives describe the dependence of U on the volume V for constant temperature, and its dependence on temperature T at constant volume, respectively. The constant parameter is indicated by the index.

According to the Boyle-Mariotte law, for any given temperature T, the product of the pressure p and V is constant (Section 1.3.3). If we change either p or V at a constant temperature, then V or p will change to maintain pV constant. However, if we change the temperature T, then the product of p and V is changed by the same factor. We have noted before that the unit of the product nRT and pV is the energy unit J (Section 1.2). Altogether, we can thus conclude that the internal energy of an ideal gas depends on its temperature, but not on its pressure or volume:

$$U \propto nRT \qquad \text{eq. 1.20}$$

The relationship in eq. 1.20 is also called *Joule's law*. As a consequence of Joule's law, the internal energy of the gas remains constant during the isothermal expansion in the first step:

$$\mathrm{d}U_{\mathrm{T}_1,\mathrm{V}_1\rightarrow\mathrm{T}_1,\mathrm{V}_2} = 0 \qquad \text{eq. 1.21}$$

and

$$\left(\frac{\partial U}{\partial V}\right)_{\mathrm{T}} = 0 \qquad \text{eq. 1.22}$$

We now define the temperature dependence of the internal energy U at constant volume as

$$\left(\frac{\partial U}{\partial T}\right)_{\mathrm{V}} = C_{\mathrm{V}} \qquad \text{eq. 1.23}$$

C_{V} is the *heat capacity* at constant volume (in J K^{-1}). For the second term in eq. 1.19, the change in internal energy upon the isochoric temperature change, we can then write

$$\mathrm{d}U_{\mathrm{T}_1,\mathrm{V}_2\rightarrow\mathrm{T}_2,\mathrm{V}_2} = C_{\mathrm{V}}\,\mathrm{d}T \qquad \text{eq. 1.24}$$

We can now combine the changes in internal energy in the two steps (eq. 1.21 and eq. 1.24) to calculate the total change in internal energy of the gas during the adiabatic expansion. As there is no heat exchange during an adiabatic expansion, the change in internal energy equals the expansion work:

$$\mathrm{d}U = w = \int_{T_1}^{T_2} C_{\mathrm{V}}\,\mathrm{d}T = C_{\mathrm{V}}\left(T_2 - T_1\right) \qquad \text{eq. 1.25}$$

Eq. 1.25 is in agreement with the system-centered sign convention: work upon expansion ($T_2 < T_1$) will be negative, meaning the system (gas) performs work on its surroundings. During a compression ($T_2 > T_1$), the work will be positive, the surroundings perform work on the system.

Table 1.1 summarizes the work associated with the state changes that we have discussed so far.

TABLE 1.1:
Work associated with reversible, irreversible, and adiabatic expansion and compression.

	irreversible	reversible	adiabatic
work w	$w = -p_{ex}\Delta V$	$w = -nRT\ln\frac{V_2}{V_1}$	$w = C_V(T_2 - T_1)$
expansion	$\Delta V > 0 \rightarrow w < 0$	$V_2 > V_1 \rightarrow w < 0$	$T_2 < T_1 \rightarrow w < 0$
compression	$\Delta V < 0 \rightarrow w > 0$	$V_2 < V_1 \rightarrow w > 0$	$T_2 > T_1 \rightarrow w > 0$

p_{ex}: external (final) pressure during an irreversible process, V_1/T_1: initial volume/temperature, V_2/T_2: final volume/temperature, C_V: heat capacity at constant volume.

We learnt earlier that the product of pressure and volume is constant for an ideal gas during isothermal processes (law of Boyle-Mariotte). If we know the initial pressure and volume of the gas before an isothermal state change and either the pressure or the volume after the change, we can therefore calculate the unknown volume or pressure from eq. 1.26:

$$p_1V_1 = p_2V_2 \qquad \text{eq. 1.26}$$

During an adiabatic process, the temperature changes, and so the product of pressure and volume before and after the process will be different. To derive a relationship between pressure and volume before and after an adiabatic process, we start from the change in internal energy dU. In the absence of other types of work and of heat exchange, the change in internal energy will equal the expansion work:

$$\mathrm{d}U = w = -p\,\mathrm{d}V \qquad \text{eq. 1.27}$$

We have derived an expression for dU during an adiabatic process in terms of the heat capacity C_V at constant volume (eq. 1.24), and we can now equate these two expressions:

$$-p\,\mathrm{d}V = C_V\,\mathrm{d}T \qquad \text{eq. 1.28}$$

Substituting for p the expression nRT/V from the ideal gas equation and separating the variables then yields

$$-\frac{nR}{V}\mathrm{d}V = \frac{C_V}{T}\mathrm{d}T \qquad \text{eq. 1.29}$$

which we can integrate from T_1 to T_2 and V_1 to V_2 to

$$-nR\ln\frac{V_2}{V_1} = C_V\ln\frac{T_2}{T_1} \qquad \text{eq. 1.30}$$

Eq. 1.30 can be rearranged to

$$\ln\frac{T_2}{T_1} = -\frac{nR}{C_V}\ln\frac{V_2}{V_1} \qquad \text{eq. 1.31}$$

which equals

$$\frac{T_2}{T_1} = \left(\frac{V_2}{V_1}\right)^{-\frac{nR}{C_V}} \qquad \text{eq. 1.32}$$

For an ideal gas, nR is the difference between the heat capacity at constant pressure, C_p (Section 1.4), and the heat capacity at constant volume C_V. Using this relationship, we can re-write eq. 1.32 as

$$\frac{T_2}{T_1} = \left(\frac{V_2}{V_1}\right)^{-\frac{C_p - C_V}{C_V}} \qquad \text{eq. 1.33}$$

Now we define the ratio of C_p/C_V as γ, and obtain

$$\frac{T_2}{T_1} = \left(\frac{V_2}{V_1}\right)^{1-\gamma} \qquad \text{eq. 1.34}$$

The gas follows the ideal gas law before and after the change in state, and we can therefore express the temperatures T_1 and T_2 by the product of p and V as

$$\frac{p_2V_2}{p_1V_1} = \left(\frac{V_2}{V_1}\right)^{1-\gamma} \qquad \text{eq. 1.35}$$

Rearrangement of eq. 1.35 finally leads us to the relationship

$$p_1V_1^{\gamma} = p_2V_2^{\gamma} \qquad \text{eq. 1.36}$$

In contrast to isothermal processes, where the product pV is constant, we see from eq. 1.36 that during adiabatic processes the product pV^{γ} is constant. Note that this does not mean that the gas does not follow the ideal gas law during the adiabatic process! In fact, we have extensively used the ideal gas law in deriving eq. 1.36. The relationship in eq. 1.36 is a consequence of the change in temperature during the adiabatic process. In analogy to the definition of an isotherm, *i.e.* the p–V curve during an isothermal process, the corresponding p–V curve for adiabatic processes is called the *adiabate*. Because of the exponent γ, the adiabate is steeper than the isotherm (Figure 1.14).

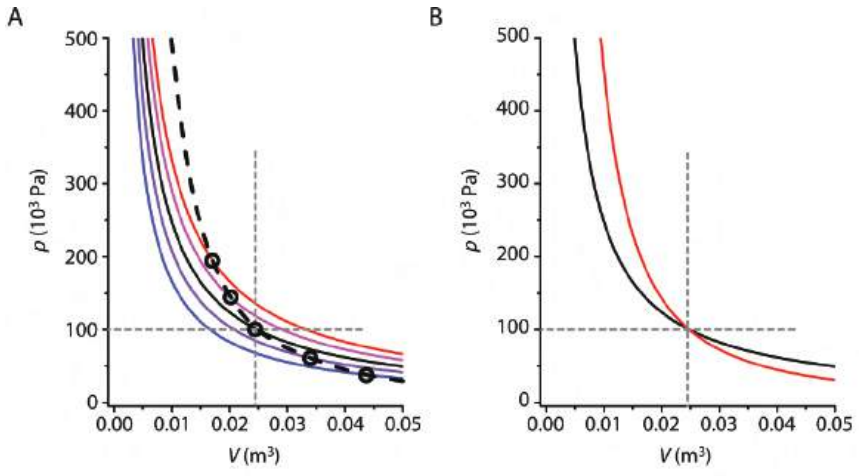


FIGURE 1.14: Comparison of isothermal and adiabatic processes. A: For isothermal processes, the product pV is constant, giving rise to a set of isotherms as described by the Boyle-Mariotte law (blue to red for increasing temperatures T). During adiabatic processes, no heat is exchanged. Therefore, adiabatic expansion leads to a decrease in temperature, and an adiabatic compression leads to an increase in temperature. This results in a crossing of the isotherms and gives rise to a steeper p–V curve, the adiabate (black dotted line). For adiabatic processes, the product pV^{γ} is constant. B: Direct comparison of isotherm (black) and adiabate (red).

1.3.5 Thermodynamic cycles: back and forth or round and round

We will now take a look at *thermodynamic cycles*. Thermodynamic cycles consist of a sequence of processes, 1→2→…→n→1, starting from an initial state (1), and at the end returning to this initial state (1). In biochemistry we frequently find cyclic processes in enzyme catalysis where an enzyme populates a sequence of different states and returns to its original state at the end of the reaction sequence (see Chapters 13 and 14; Box 1.4). Membrane transport is also based on transporters cycling through different states (Section 4.3, Box 1.4).

The simplest form of a thermodynamic cycle is a coupling of two opposite changes, back and forth, such as the expansion of an ideal gas, followed by the compression to its original state. We first look at an irreversible expansion of an ideal gas at p_1,V_1 to p_2,V_2, followed by an irreversible compression back to the starting state with p_1,V_1. The volume change ΔV is V_2–V_1 for the expansion and V_1–V_2 for the compression, and thus equal in size, but opposite in sign. The work w_{tot} over the complete process is then simply the sum of the work involved in the two steps:

$$w_{tot} = w_{exp} + w_{comp} = -p_2\Delta V + p_1\Delta V \qquad \text{eq. 1.37}$$

As $p_1 > p_2$, more work has to be performed on the system during compression at p_1 compared to the work the system has performed on its surroundings upon expansion at p_2. As a consequence, we obtain:

$$w_{tot} > 0 \qquad \text{eq. 1.38}$$

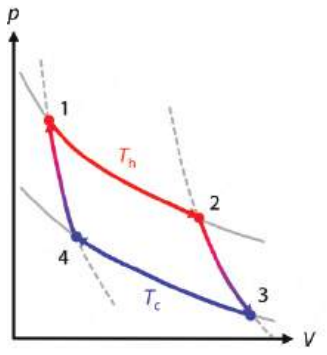
If we now perform the sequence of expansion and compression to the initial state under reversible conditions, the pressure will be equal to the gas pressure according to the ideal gas law throughout the process. Under reversible conditions, we therefore obtain for w_{tot}:

$$w_{tot} = w_{exp} + w_{comp} = -nRT\ln\frac{V_2}{V_1} - nRT\ln\frac{V_1}{V_2} \qquad \text{eq. 1.39}$$

By inverting the argument of the logarithm for the compression, we change the sign of the second term, and see that now the work performed by the system during expansion equals the work that has to be performed on the system during compression of the gas. As a consequence, w_{tot} is zero.

1.3.5.1 The Carnot process

We can now think of a cyclic process of the sequence 1→2→3→4→1, *i.e.* from an initial state 1 through intermediate states 2, 3, and 4 back to state 1. This cycle consists of four steps: (1) an isothermal expansion at the temperature T_h (1→2), (2) an adiabatic expansion (2→3), (3) an isothermal compression at the temperature T_c (3→4), and (4) an adiabatic compression (4→1). The temperatures T_h and T_c are maintained constant by coupling the system to two heat reservoirs, a hot reservoir at T_h during the isothermal expansion, and a cold reservoir at T_c during the isothermal compression. Such a cyclic process is called a *Carnot cycle* (Figure 1.15).



$1 \rightarrow 2$:
$w_{1\rightarrow2} = -nRT_h \ln(V_2/V_1)$
$q_{1\rightarrow2} = -w_{1\rightarrow2}$

$2 \rightarrow 3$:
$w_{2\rightarrow3} = C_v(T_c - T_h)$

$3 \rightarrow 4$:
$w_{3\rightarrow4} = -nRT_c \ln(V_1/V_2)$
$q_{3\rightarrow4} = -w_{3\rightarrow4}$

$4 \rightarrow 1$:
$w_{4\rightarrow1} = C_v(T_h - T_c)$

FIGURE 1.15: The Carnot cycle. The Carnot process is a cyclic process in which an ideal gas is first expanded under isothermal conditions at the temperature T_h, and then further expanded under adiabatic conditions, followed by an isothermal compression at the temperature T_c, and an adiabatic compression back to the starting state. The exchanged heat and work in each step is summarized on the right (see text).

In the Carnot process, the initial state of the system is equal to its final state after completion of one cycle. As a consequence, the internal energy U (and other state functions) is constant over the complete cycle. The associated work with the isothermal processes 1→2 (expansion at T_h) and 3→4 (compression at T_c) is

$$w_{\text{isothermal}} = w_{1\rightarrow2} + w_{3\rightarrow4} = -nRT_h \ln\frac{V_2}{V_1} - nRT_c \ln\frac{V_4}{V_3} \qquad \text{eq. 1.40}$$

From eq. 1.34 for adiabatic changes, we can derive a relationship between T_h and T_c and V_2 and V_3 for the adiabatic expansion, and between T_c and T_h and V_4 and V_1 for the adiabatic compression. By solving both relationships for T_h/T_c and equating them, we arrive at

$$\frac{V_1}{V_2} = \frac{V_4}{V_3} \qquad \text{eq. 1.41}$$

Using eq. 1.41, we can eliminate V_3 and V_4 from eq. 1.40, and obtain for the work associated with the two isothermal processes

$$w_{\text{isothermal}} = w_{1\rightarrow2} + w_{3\rightarrow4} = -nRT_h \ln\frac{V_2}{V_1} + nRT_c \ln\frac{V_2}{V_1} = -nR(T_h - T_c)\ln\frac{V_2}{V_1} \qquad \text{eq. 1.42}$$

The work for the two adiabatic processes 2→3 (expansion) and 4→1 (compression) is

$$w_{\text{adiabatic}} = w_{2\rightarrow3} + w_{4\rightarrow1} = C_V(T_c - T_h) + C_V(T_h - T_c) = 0 \qquad \text{eq. 1.43}$$

Finally, the work over the complete cycle is the sum of the work for all individual steps:

$$w_{\text{tot}} = w_{\text{isothermal}} + w_{\text{adiabatic}} = -nR(T_h - T_c)\ln\frac{V_2}{V_1} \qquad \text{eq. 1.44}$$

Because $T_h > T_c$ and $V_2 > V_1$, w_{tot} is negative, and over the complete cycle the system thus performs the work w_{tot} on the surroundings!

From the work involved in each step of the Carnot cycle, we can now calculate the heat exchanged. Heat is only exchanged during the isothermal steps, and the exchanged heat again has the same magnitude as the work performed, but the opposite sign:

$$q_{\text{isothermal}} = -(w_{1\rightarrow2} + w_{3\rightarrow4}) = nR(T_h - T_c)\ln\frac{V_2}{V_1} \qquad \text{eq. 1.45}$$

Again, because $T_h > T_c$ and $V_2 > V_1$, the overall heat exchange is positive, and the system absorbs heat during one cycle. In sum, during each complete Carnot cycle, the system takes up the heat q_h from the hot reservoir at the temperature T_h, performs the work w_{tot} on the surroundings, and transfers the heat q_c to the cold reservoir at the temperature T_c (Figure 1.16). Thus, the heat taken up by the system is partially converted into work.

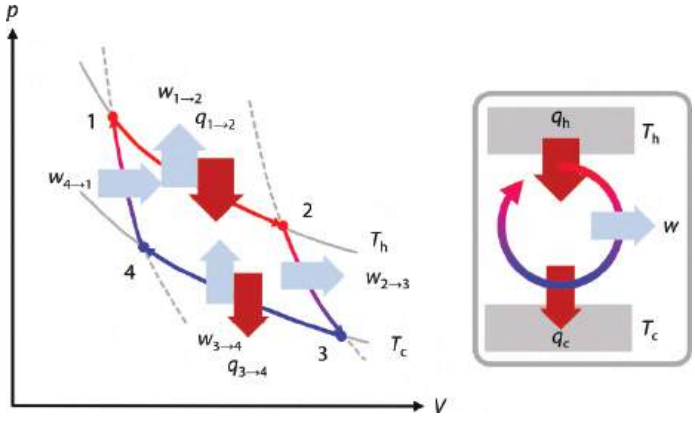


FIGURE 1.16: Heat and work during the Carnot process. In the isothermal expansion (1→2), the system performs work on the surroundings ($w < 0$), and takes up heat ($q > 0$). During the adiabatic expansion (2→3), work is performed on the surroundings ($w < 0$), but no heat is exchanged. In the isothermal compression (3→4), the surroundings perform work on the system ($w > 0$), and heat is transferred from the system to the surroundings ($q < 0$). In the adiabatic compression (4→1), work is performed on the system ($w > 0$) without heat exchange. The work associated with the two adiabatic processes is equal, but of different sign, and cancels. The work performed on the surroundings upon isothermal expansion at T_h is larger than the work performed on the system during the isothermal compression at T_c. Thus, over the complete process the system performs work on the surroundings. The heat taken up by the system from the hot reservoir during the isothermal expansion at T_h is larger than the heat transferred from the system to the cold reservoir during the isothermal compression at T_c. Overall, the system thus takes up heat from the hot reservoir, converts part of it into work that is performed on the surroundings, and releases the remaining heat to the cold reservoir.

The *degree of efficiency* η of the Carnot cycle in Figure 1.15 and Figure 1.16 can be described as the performed work $-w_{tot}$ relative to the heat absorbed at T_h, q_h (eq. 1.46). Note that by taking the negative work $-w_{tot}$, we now change the perspective from the system to a perspective centered on the surroundings: w_{tot} is the work from the system-centered perspective, and negative values of w_{tot} mean the system performs work on the surroundings. By inverting the sign to $-w_{tot}$ we convert the work into a positive number, the work that can be used for processes in the surroundings.

$$\eta_{Carnot} = \frac{-w_{tot}}{q_h} = \frac{nR(T_h - T_c)\ln\frac{V_2}{V_1}}{nRT_h \ln\frac{V_2}{V_1}} = \frac{T_h - T_c}{T_h} \qquad \text{eq. 1.46}$$

From eq. 1.46, we see that the degree of efficiency increases with increasing differences of T_h and T_c. It would reach a limiting value of unity (*i.e.* 100% efficiency) only for $T_c = 0$ K, the absolute zero point. We have seen before (Section 1.3.3) that the absolute zero point cannot be reached. As a consequence, η will always remain < 1. Thus, from eq. 1.46 we can conclude that, for thermodynamic reasons, heat cannot be completely converted into work! This statement is one form of the second law of thermodynamics. We will arrive at an alternative form of the second law, the Clausius inequality, in Section 1.5.2.

The Carnot cycle in clockwise direction, following the sequence 1→2→3→4→1 that partially converts heat into work (Figure 1.15 and Figure 1.16) is called a *forward Carnot cycle*. Such a process is the basis of a power heat pump such as a steam engine. We can also reverse the sequence, and

run the process counter-clockwise, starting with an adiabatic expansion (1→4), followed by an isothermal expansion (4→3), an adiabatic compression (3→2), and an isothermal compression (2→1) (Figure 1.17). In this case, the Carnot process will transport heat from the cold reservoir (T_c) to the hot reservoir (T_h). We will see in Section 1.5.1 (Figure 1.19) that the transfer of heat from cold to hot is against the spontaneous dissipation of heat, and thus an energetically unfavorable process. To drive this unfavorable process, work has to be performed on the system (Figure 1.17). Such a *reverse Carnot cycle* is the basic principle behind refrigerators, which need electric power to transport heat from their colder interior to the warmer room, or for heat pumps that take up heat from the cooler air or from deep in the ground, and pump it into the warmer living room, again at the expense of electric power. We will re-encounter this principle when we discuss transport across biological membranes (Section 4.3; Box 1.4).

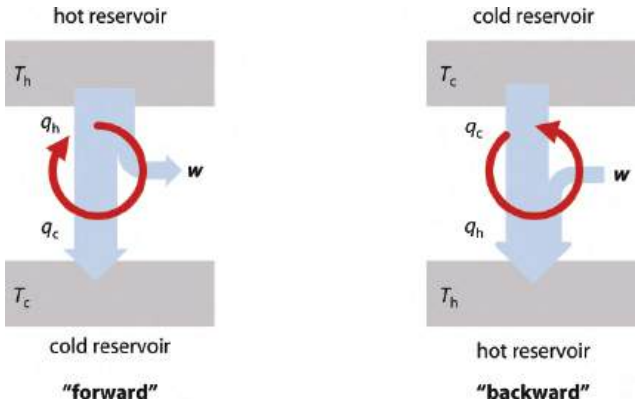


FIGURE 1.17: A Carnot cycle in forward and reverse mode. In forward mode, following the sequence 1→2→3→4→1 (see Figure 1.15), the Carnot process converts part of the heat *q* it takes up from the hot reservoir into work *w* performed on the surroundings. The rest of the heat taken up is released to the cold reservoir. This process is the principle behind steam engines. In reverse mode, following the sequence 1→4→3→2→1, the Carnot process takes up energy from a cold reservoir, and, together with work taken up from the surroundings, transfers a larger amount of heat to a hot reservoir. This reverse mode is the basis for the function of refrigerators and of heat pumps.

BOX 1.4: CYCLIC PROCESSES IN BIOCHEMISTRY.

In biochemistry, reactions typically occur at isothermal and isobaric conditions; nevertheless, we find many reaction cycles that bear resemblance to the Carnot cycle. A number of enzymes convert one form of energy into a different form. An example that we will treat in Section 4.3 is the class of ATP-driven transporters that use the energy of ATP hydrolysis (chemical energy) to establish concentration gradients (osmotic energy). The transport cycle is similar to a reverse Carnot cycle: the transporter takes up cargo from a low concentration side, which corresponds to the process of taking up heat from a cold reservoir. The cargo is transported across a membrane and released on the distant side at a high concentration, which corresponds to the release of heat into a warm reservoir. Other transporters can use the energy of light for active transport. Further examples include motor proteins that couple the energy of ATP hydrolysis (chemical energy) to generate movement (mechanical energy), and enzymes that use the energy of ATP hydrolysis to alter the structure of their substrate, such as topoisomerases that can introduce torsional stress into DNA in an ATP-dependent reaction. During their catalytic cycles, these enzymes undergo conformational changes that contribute to coupling ATP hydrolysis to movement or to the catalyzed reaction. At the end of the catalytic cycle, the chemical energy has been converted into a different type of energy, and the enzyme is regenerated in its original state and can undergo further catalytic cycles.

(*Continued*)

Cyclic processes can also be exploited to generate heat. Bumble bees use a substrate cycle to heat their bodies: they phosphorylate fructose-6-phosphate to fructose-1,6-bisphosphate, using ATP as a phosphoryl group donor. Fructose-1,6-bisphosphate is then dephosphorylated, leading to regeneration of fructose-6-phosphate. During each cycle, one ATP molecule is hydrolyzed. The energy of ATP hydrolysis is converted into heat in this process. The resulting increase in body temperature warms up flight muscles to a temperature sufficient to generate force. This process is called *thermogenesis*. The heat generation through the substrate cycle enables bumble bees to fly at low temperatures, whereas honey bees have to wait until the outside temperature warms them up.

1.3.6 The temperature dependence of the internal energy U

If we consider an ideal gas and allow only expansion work, the change $\mathrm{d}U$ in the internal energy U is described by

$$\mathrm{d}U = -p\,\mathrm{d}V + \delta q \qquad \text{eq. 1.47}$$

For isochoric processes (V = const., $\mathrm{d}V = 0$), eq. 1.47 simplifies to

$$\mathrm{d}U = \delta q_\mathrm{V} \qquad \text{eq. 1.48}$$

where the index V denotes the isochoric conditions. Thus, by limiting our considerations to changes with constant volume, we have converted a non-state function, the heat q, into a state function, the internal energy U (under certain conditions, *i.e.* for isochoric processes). For processes at constant volume, the change in internal energy thus equals the amount of heat exchanged. This heat will affect the temperature of the system. We have defined the dependence of U on the temperature at constant volume as the heat capacity C_V at constant volume (eq. 1.23). Now we can relate the heat capacity C_V to the heat exchanged in an isochoric process:

$$\left(\frac{\partial U}{\partial T}\right)_\mathrm{V} = C_\mathrm{V} = \frac{\delta q_\mathrm{V}}{\mathrm{d}T} \qquad \text{eq. 1.49}$$

We have derived the relationship between $\mathrm{d}U$ and $\mathrm{d}T$ before (eq. 1.24), and can now directly derive the temperature dependence of U (at constant volume) by integrating eq. 1.24 from T_0 to T:

$$\int_{U(T_0)}^{U(T)} \mathrm{d}U = \int_{T_0}^{T} C_\mathrm{V}\,\mathrm{d}T \qquad \text{eq. 1.50}$$

We assume that C_V is independent of the temperature in the temperature interval from T_0 to T, and obtain the temperature dependence of the internal energy U by evaluating the integrals as

$$U(T) - U(T_0) = C_\mathrm{V}(T - T_0) \qquad \text{eq. 1.51}$$

or

$$U(T) = U(T_0) + C_\mathrm{V}(T - T_0) \qquad \text{eq. 1.52}$$

Thus, if we know the internal energy of our system at some reference temperature T_0, and we know the heat capacity C_V (at constant volume), we can calculate the internal energy U of the system at any other temperature T, provided that C_V is temperature-independent.

U is a state function, hence changes in *U* that are associated with any process can be described as

$$\Delta U = U_{\text{final}} - U_{\text{initial}} \qquad \text{eq. 1.53}$$

If we now consider the temperature dependence of ΔU, $\mathrm{d}/\mathrm{d}T\,(\Delta U)$, we obtain

$$\frac{\mathrm{d}(\Delta U)}{\mathrm{d}T} = \frac{\mathrm{d}(U_{\text{final}})}{\mathrm{d}T} - \frac{\mathrm{d}(U_{\text{initial}})}{\mathrm{d}T} = C_{\mathrm{V,final}} - C_{\mathrm{V,initial}} \qquad \text{eq. 1.54}$$

Eq. 1.54 relates the change in internal energy during an isochoric process to the change in heat capacity ΔC_{V} between final and initial states:

$$\frac{\mathrm{d}(\Delta U)}{\mathrm{d}T} = \Delta C_{\mathrm{V}} \qquad \text{eq. 1.55}$$

By integrating eq. 1.55, we obtain ΔU, the change in the internal energy during this process, as a function of temperature *T* and the change in heat capacity ΔC_{V} associated with the process:

$$\Delta U(T) = \Delta U(T_0) + \Delta C_{\mathrm{V}}(T - T_0) \qquad \text{eq. 1.56}$$

The internal energy *U* is a useful state function to describe processes at constant volume, where it corresponds to the exchanged heat. Biological processes typically take place at constant pressure, not at constant volume, but we will see in the following Section that we can apply the principles we have learnt from considerations of the internal energy directly to isobaric processes.

1.4 THE ENTHALPY *H*

We have seen in Section 1.3 that the change in internal energy *U* corresponds to the heat exchanged during an isochoric process (eq. 1.48). By analogy, we can also calculate the heat exchanged under isobaric conditions. First, we rearrange eq. 1.47 to

$$\delta q = \mathrm{d}U + p\,\mathrm{d}V \qquad \text{eq. 1.57}$$

We now define a new state function (we will see in a moment why this state function is useful), the *enthalpy H*, as

$$H = U + pV \qquad \text{eq. 1.58}$$

Complete differentiation of *H* yields

$$\mathrm{d}H = \mathrm{d}U + p\,\mathrm{d}V + V\,\mathrm{d}p \qquad \text{eq. 1.59}$$

For isobaric processes (p = const., $\mathrm{d}p = 0$), eq. 1.59 simplifies to

$$\mathrm{d}H = \mathrm{d}U + p\,\mathrm{d}V \qquad \text{eq. 1.60}$$

If we now compare eq. 1.60 and eq. 1.57, we see that the right-hand sides are equal. Consequently, $\mathrm{d}H$ equals the heat exchanged at constant pressure:

$$\mathrm{d}H = \delta q_{\mathrm{p}} \qquad \text{eq. 1.61}$$

Changes dH in the enthalpy H equal the amount of heat transferred during isobaric processes. Thus, the enthalpy H at constant pressure is an equivalent state function to the internal energy U at constant volume.

We can extend the analogy of dH (for isobaric processes) and dU (for isochoric processes), and define a heat capacity C_p at constant pressure as

$$\left(\frac{\partial H}{\partial T}\right)_p = C_p \qquad \text{eq. 1.62}$$

From eq. 1.62, we obtain

$$\mathrm{d}H = C_p\,\mathrm{d}T \qquad \text{eq. 1.63}$$

Separation of the variables and integration from T_0 to T yields

$$H(T) = H(T_0) + C_p(T - T_0) \qquad \text{eq. 1.64}$$

which allows us to calculate the enthalpy H for a system at any temperature T, if we know H at a reference temperature T_0, $H(T_0)$, and C_p. Again, because H is a state function, we can calculate changes in enthalpy associated with any given process simply by subtracting the enthalpies of the initial and final states:

$$\Delta H = H_{\text{final}} - H_{\text{initial}} \qquad \text{eq. 1.65}$$

For changes in state, such as the expansion of a gas, H_{initial} and H_{final} are the enthalpies of the gas before and after the expansion. For a chemical reaction, H_{initial} is the enthalpy of the reactants, H_{final} is the enthalpy of the products, and ΔH_r is the change in enthalpy of the reaction, *i.e.* the heat exchanged during the reaction at constant pressure:

$$\Delta H_r = H_{\text{products}} - H_{\text{reactants}} \qquad \text{eq. 1.66}$$

If we now consider the temperature dependence of ΔH, d/dT (ΔH), eq. 1.65 becomes

$$\frac{\mathrm{d}(\Delta H)}{\mathrm{d}T} = \frac{\mathrm{d}(H_{\text{final}})}{\mathrm{d}T} - \frac{\mathrm{d}(H_{\text{initial}})}{\mathrm{d}T} = C_{p,\text{final}} - C_{p,\text{initial}} \qquad \text{eq. 1.67}$$

or

$$\frac{\mathrm{d}(\Delta H)}{\mathrm{d}T} = \Delta C_p \qquad \text{eq. 1.68}$$

The relationship between changes in enthalpy and heat capacity according to eq. 1.68 is called *Kirchhoff's law*. By integrating this equation, the temperature dependence of ΔH as a function of the change in heat capacity is obtained:

$$\Delta H(T) = \Delta H(T_0) + \Delta C_p(T - T_0) \qquad \text{eq. 1.69}$$

Thus, the change in enthalpy ΔH is temperature-dependent. The temperature dependence of ΔH is a consequence of the different temperature dependencies of H for the final and initial states involved in the process. For chemical reactions, the enthalpy change ΔH_r depends on the temperature because of the different temperature dependence of the enthalpies of the reactants and the products ($C_{p,\text{reactants}} \neq C_{p,\text{products}}$, $\Delta C_p \neq 0$; Figure 1.18).

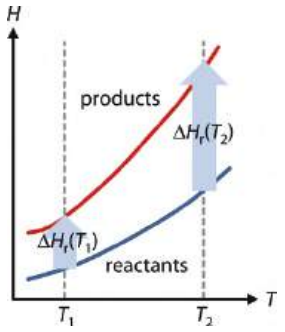


FIGURE 1.18: Temperature dependence of enthalpy and enthalpy changes. The enthalpies of the reactants and of the products depend on the temperature. In general, reactants and products will have different heat capacities, and thus a different dependence of their enthalpies on temperature. As a consequence, the change in enthalpy during the reaction, ΔH_r, the difference between $H_{products}$ and $H_{reactants}$, also depends on temperature (arrows).

We can exploit the property of the enthalpy H as a state function to calculate the enthalpy change for a process that is difficult to examine experimentally. To do so, we have to relate this process to others whose enthalpy changes are known, and use the *Hess law* of heat summation (Box 1.5).

BOX 1.5: HESS LAW OF HEAT SUMMATION.

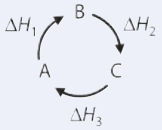
Because H is a state function, we can always calculate changes in H from the difference $\Delta H = H_{final} - H_{initial}$. For a chemical reaction, this enthalpy change is $\Delta H_r = H_{products} - H_{reactants}$. We can also use the property of H as a state function to calculate the enthalpy change for a process that is difficult to examine experimentally by relating this process to others whose enthalpy changes are known. In the simplest case, we can calculate the overall change in ΔH for a sequence of reactions (Scheme 1.1) by adding up the ΔH values for each individual step,

$$A \xrightarrow{\Delta H_1} B \xrightarrow{\Delta H_2} C \xrightarrow{\Delta H_3} \ldots \xrightarrow{\Delta H_{25}} Z \qquad \text{scheme 1.1}$$

and

$$\Delta H_{tot} = \sum_{i=1}^{25} \Delta H_i \qquad \text{eq. 1.70}$$

We can also construct cyclic schemes, *e.g.* with reactions from A to B to C and back to A (Scheme 1.2):



scheme 1.2

The change in enthalpy in one cycle (from A to A) has to be zero, because the initial and final states are identical. If we know all ΔH values except one, *e.g.* ΔH_2 and ΔH_3, we can calculate the missing ΔH_1 value as

$$\Delta H_1 = \Delta H_{tot} - \Delta H_2 - \Delta H_3 = -\Delta H_2 - \Delta H_3 \qquad \text{eq. 1.71}$$

The additivity of heat and changes in ΔH is called the *Hess law* of heat summation. We can transfer this same principle to any other state function, *e.g.* to the changes in entropy ΔS (Section 1.5) or free energy ΔG (Section 1.6).

Isobaric processes with $\Delta H > 0$ are associated with an uptake of heat ($\delta q_p > 0$), and are called *endothermic*. Processes with $\Delta H < 0$ are accompanied by a release of heat ($\delta q_p < 0$), and are called *exothermic*. Most organisms live at constant pressure, and ΔH and ΔC_p (rather than ΔU and ΔC_v) are therefore important parameters of life.

1.5 THE ENTROPY *S* AND THE SECOND LAW OF THERMODYNAMICS

1.5.1 PREDICTING SPONTANEITY OF PROCESSES: DISSIPATION OF HEAT AND MATTER

Until now we have considered the overall energy balance associated with expansion and compression of gases under different conditions, and have calculated the work and heat exchanged during these processes using the first law of thermodynamics. However, the first law of thermodynamics does not allow us to predict the spontaneous direction of a process. For example, both sodium nitrate ($NaNO_3$) and sodium hydroxide (NaOH) dissolve spontaneously in water. The dissolution of NaOH is an exothermic reaction with $\Delta H < 0$: the solution becomes warm, and heat is transferred to the surroundings. In contrast, the dissolution of $NaNO_3$ is an endothermic process, with $\Delta H > 0$. As a consequence, the solution becomes cold, and heat is drawn from the surroundings. Despite the opposite sign of ΔH, both reactions are spontaneous. We can conclude that the sign of ΔH does therefore not have any predictive value for the spontaneity of reactions. The internal energy U likewise does not indicate the direction of spontaneous processes. For example, we know from our everyday experience that we can leave a kettle with hot water on the stove, and it will slowly transfer heat to the room and cool down to room temperature. The reverse process, in which heat is spontaneously withdrawn from the room to bring water in the kettle to a boil, is never observed. More generally, bringing together a cold and a hot object will lead to heat exchange in one direction only, from the hot to the cold object, until both have the same temperature and are in thermal equilibrium. This process toward an even distribution of heat is called *dissipation* (Figure 1.19). However, the internal energy of the system, defined as the room including the kettle, or the two objects of different temperatures, does not change because heat is just moved from one element to the other within the system.

Dissipation is not only observed for heat, but also for matter: if we allow a gas to occupy a larger volume, it will spontaneously do so. The reverse process, the spontaneous retraction of the gas to only part of the available volume, is never observed (Figure 1.19). The expansion of a gas under isothermal conditions does not alter the internal energy of the gas. In general, the internal energy of a system does not change during the dissipation of matter.

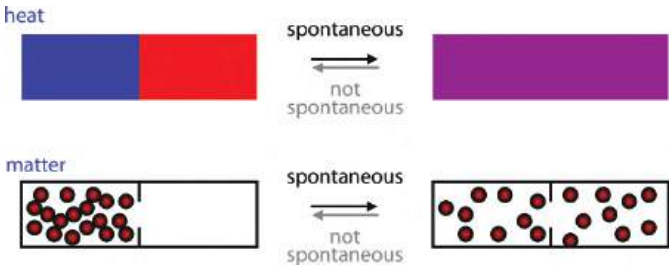


FIGURE 1.19: Dissipation of heat and matter. When we bring a hot (red) and a cold (blue) object in contact, heat will spontaneously flow from the hot to the cold object until both objects have the same temperature (thermal equilibrium, violet). The reverse process does not occur spontaneously. Similarly, when we provide a larger volume for a gas (red particles), the particles will spontaneously occupy the total available volume. The reverse processes, in which particles accumulate in a small area of the available volume, do not occur spontaneously. Both heat and matter have a tendency to distribute evenly, a process called dissipation.

We have been able to intuitively predict the direction of these dissipative processes. If we compare initial and final states, we see that the final state, with heat or matter distributed throughout the system, is a state of lower order than the initial state. How can we formulate the underlying principle, a decrease in order, in a more objective and quantitative manner? What we need is a state function that only changes in one direction, such that the sign of its change predicts if a reaction is spontaneous. In the following, we will derive this state function from heat considerations (Section 1.5.2). Later, we will see that it is connected to the degree of order in the system (Section 1.5.5).

1.5.2 Entropy and heat

To define a state function that predicts the spontaneous direction of a process, we return to the Carnot cycle we have considered in Section 1.3.5. We have calculated the heat exchanged over the complete cycle (eq. 1.45),

$$q_{\text{isothermal}} = q_{1\to2} + q_{3\to4} = nR\left(T_\text{h} - T_\text{c}\right)\ln\frac{V_2}{V_1} \qquad \text{eq. 1.45}$$

and we have seen that the overall heat exchanged is positive. This is in agreement with heat not being a state function. We have seen before that we can nevertheless convert heat into a state function under certain conditions, *i.e.* into the enthalpy *H* under isobaric conditions, or into the internal energy *U* under isochoric conditions. We will now illustrate how we can convert heat into a state function if we normalize it by the temperature. Instead of the heat we now sum the "relative heat" for one Carnot cycle, *i.e.* the heat divided by the absolute temperature at which the process takes place. We obtain

$$\frac{q_{1\to2}}{T_\text{h}} + \frac{q_{3\to4}}{T_\text{c}} = nR\ln\frac{V_2}{V_1} - nR\ln\frac{V_2}{V_1} = 0 \qquad \text{eq. 1.72}$$

While the heat does not qualify as a state function, the sum of the "relative heat" now equals zero over the cyclic process, and thus has the property of a state function! We therefore define this new state function as the entropy *S* with

$$S = \frac{q_\text{rev}}{T} \qquad \text{eq. 1.73}$$

The index "rev" denotes that this definition is derived for reversible processes. Comparing with eq. 1.72, we see that for a complete Carnot cycle consisting only of reversible processes, the change in entropy is

$$\mathrm{d}S = \frac{\delta q_\text{rev}}{T} = 0 \qquad \text{eq. 1.74}$$

The equality (eq. 1.74) only holds true for reversible processes, *i.e.* for systems in equilibrium. For irreversible processes, eq. 1.74 becomes an inequality (see eq. 1.82, Box 1.7).

We can derive a more general form of eq. 1.74 by including irreversible processes in the Carnot cycle. We know that irreversible processes, such as friction, will lead to the conversion of energy to heat, and will reduce the overall degree of efficiency η:

$$\eta_\text{irrev} < \eta_\text{Carnot,rev} \qquad \text{eq. 1.75}$$

Biological processes are often reversible, which is the reason why biomolecular machines can reach high degrees of efficiency (Box 1.6). For the Carnot cycle, we have defined the degree of efficiency in general as the ratio of work performed by the system on the surroundings, from the perspective of the surroundings ($-w_\text{tot}$), divided by the heat q_h taken up from the high temperature reservoir

(eq. 1.46). This general definition also holds for irreversible conditions. For reversible processes, we can calculate the individual heat terms, and express the degree of efficiency in terms of the temperatures T_h and T_c. We can therefore relate the degrees of efficiency under irreversible and reversible conditions by:

$$\eta_{irrev} = \frac{-w_{tot}}{q_{1\to2}} = \frac{q_{1\to2} + q_{3\to4}}{q_{1\to2}} < \eta_{Carnot} = \frac{T_h - T_c}{T_h} \qquad \text{eq. 1.76}$$

leading to the relationship

$$\frac{q_{1\to2} + q_{3\to4}}{q_{1\to2}} < \frac{T_h - T_c}{T_h} \qquad \text{eq. 1.77}$$

We can rearrange this inequality to

$$1 + \frac{q_{3\to4}}{q_{1\to2}} < 1 - \frac{T_c}{T_h} \qquad \text{eq. 1.78}$$

and

$$\frac{q_{3\to4}}{q_{1\to2}} < -\frac{T_c}{T_h} \qquad \text{eq. 1.79}$$

and arrive at

$$\frac{q_{1\to2}}{T_h} + \frac{q_{3\to4}}{T_c} < 0 \qquad \text{eq. 1.80}$$

The left-hand side of eq. 1.80 is the sum of all q/T terms, or the circular integral (see Section 31.6) over the Carnot cycle. The zero on the right-hand side is the overall entropy change over one Carnot cycle, which as a state function equals zero over a circular process. Eq. 1.80 can thus be re-written as

$$\oint \frac{\delta q}{T} < \oint dS \qquad \text{eq. 1.81}$$

The circle in eq. 1.81 denotes integration over a circular process (see Section 31.6). By comparing the arguments of the integrals only, we can conclude that for the entire cycle:

$$\frac{\delta q}{T} < dS \qquad \text{eq. 1.82}$$

During a Carnot cycle with irreversible elements, the entropy increase is thus larger than the transferred heat divided by the temperature. Irreversible processes are associated with *entropy production* in the surroundings. This process enables the formation of ordered systems in biology, away from equilibrium (Box 1.7). For a reversible Carnot cycle, the inequality in eq. 1.82 becomes an equality. Eq. 1.82 is an alternative formulation of the *second law of thermodynamics*. It is a simple statement with important consequences: the reason why heat cannot be converted completely into work.

For isolated systems, δq is zero. In this case, eq. 1.82 simplifies to the relationship

$$dS \geq 0 \qquad \text{eq. 1.83}$$

which is the *Clausius inequality*. From eq. 1.83, we see that the entropy is the state function with the predictive value for spontaneous processes that we were looking for: in isolated systems, a process is spontaneous when the entropy increases. Processes with constant entropy are in equilibrium, and processes with a decrease in entropy do not occur spontaneously.

BOX 1.6: REVERSIBILITY AND THE DEGREE OF EFFICIENCY.

We have learnt in Section 1.3.4.3 that we can have a system perform most work under reversible conditions. In a car engine, the ignition and combustion of fuel causes a rapid and irreversible adiabatic expansion. This irreversibility is inevitable if we want to convert the energy of combustibles into work, yet it is the reason for the rather poor degrees of efficiency of these machines. Biological machines, on the other hand, achieve very high degrees of efficiencies because they perform their work under reversible conditions. Many of these machines use ATP as an energy source, and hydrolyze ATP to ADP and phosphate. However, these machines are true enzymes that also catalyze the reverse reaction (Section 16.2), the condensation of ADP and phosphate to ATP. The reversibility of reactions catalyzed by biological machines and motors is the key to their impressively high degrees of efficiency.

BOX 1.7: THERMODYNAMICS OF IRREVERSIBLE PROCESSES.

In reversible processes, heat is exchanged between system and surroundings. The resulting entropy increase ΔS_{system} in the system is mirrored by a decrease of equal size in the surroundings, $\Delta S_{surroundings} = -\Delta S_{system}$ (see Section 1.6). This phenomenon is termed *entropy transport*.

$$\Delta S_{system} + \Delta S_{surroundings} = 0 \qquad \text{eq. 1.84}$$

We have seen that the situation is different for irreversible processes. Again, entropy is exchanged between the system and its surroundings. However, the entropy increase associated with an irreversible process is larger than the heat transferred from the system to the surroundings divided by the temperature (eq. 1.82). Thus, an irreversible process within the system must lead to the production of entropy in the surroundings: eq. 1.84 becomes an inequality.

$$\Delta S_{system} + \Delta S_{surroundings} = \Delta S_{prod} > 0 \qquad \text{eq. 1.85}$$

The sum in eq. 1.85 now corresponds to the entropy ΔS_{prod} that is produced in this process (see Section 1.6). ΔS_{prod} is a measure for the degree of irreversibility.

For the analysis of irreversible processes, the change in state functions with time is a key parameter. Here, we limit our considerations to changes in entropy, specifically to the entropy ΔS_{prod} produced in an irreversible process. We define the *entropy production* $\dot{\sigma}$ as the change in ΔS_{prod} with time:

$$\dot{\sigma} = \frac{d\left(\Delta S_{prod}\right)}{dt} \qquad \text{eq. 1.86}$$

Formulation of $\dot{\sigma}$ is the starting point for the analysis of irreversible processes. We now define the entropy σ of a system as the difference between its entropy S in a state away from equilibrium, and its entropy in equilibrium, S_0:

$$\sigma = S - S_0 \qquad \text{eq. 1.87}$$

(*Continued*)

In equilibrium, with $S = S_0$, σ becomes zero. We know that in equilibrium, the entropy is maximized. Away from equilibrium, we therefore must have

$$\sigma < 0 \qquad \text{eq. 1.88}$$

As a consequence, the entropy production $\dot{\sigma}$ must be positive ($\dot{\sigma} > 0$) for a system away from equilibrium, decrease on its way toward equilibrium, and become zero when the equilibrium has been reached. The *theorem of minimal entropy production* states that systems develop in the direction where the entropy production becomes minimal. Thereby, they reach a stable state in which the parameters describing the system remain constant. The state with minimal entropy production is also termed *steady state*. We will re-encounter steady states in enzyme kinetics (Chapter 13). The steady state is the form of open systems that is equivalent to the equilibrium in closed systems. However, it differs from the equilibrium state: while the equilibrium state is a disordered state with maximized entropy, the steady state is an ordered state with low entropy. The thermodynamic description of irreversible processes depends on how far away the steady state is from the thermodynamic equilibrium. If the steady state is close to the thermodynamic equilibrium, the process is a linear process, if it is far away from equilibrium, the process is non-linear. This distinction is based on the expansion of the entropy production in a Taylor series (see Section 31.14.1): a linear process can be described with only the linear terms of the Taylor series, whereas the description of non-linear processes requires the non-linear terms as well.

Living organisms are examples of ordered, low-entropy systems in a steady state, away from thermodynamic equilibrium. Organisms are open systems: the steady state is maintained by a constant flux of energy and matter from the surroundings to the organism and from the organisms to its surroundings. Thermodynamically, the entropically unfavorable generation of order in the form of an organism becomes possible by entropy production in the surroundings. Once the organism dies, the steady state breaks down, and the system will strive to maximize entropy and reach thermodynamic equilibrium.

Non-equilibrium states are a recurrent principle in biology. Efficient regulation is only possible for processes that are far away from equilibrium. Molecular motors can bias Brownian motion into directed motion if the system is far away from equilibrium (Section 4.4).

1.5.3 Temperature dependence of the entropy

For processes at constant pressure, we can express changes in entropy as

$$\mathrm{d}S = \frac{\delta q_{\mathrm{rev}}}{T} = \frac{C_{\mathrm{p}}\,\mathrm{d}T}{T} \qquad \text{eq. 1.89}$$

Integration of eq. 1.89 from T_0 to T yields

$$S(T) - S(T_0) = C_{\mathrm{p}} \ln \frac{T}{T_0} \qquad \text{eq. 1.90}$$

or

$$S(T) = S(T_0) + C_{\mathrm{p}} \ln \frac{T}{T_0} \qquad \text{eq. 1.91}$$

By analogy, we obtain for the change in entropy during a reaction, $\Delta S_{\mathrm{r}}(T)$

$$\Delta S_r(T) = \Delta S_r(T_0) + \Delta C_p \ln \frac{T}{T_0} \qquad \text{eq. 1.92}$$

where ΔC_p is the change in heat capacity during the reaction. The entropy change is temperature-dependent because of the different temperature dependencies of the entropies of reactants and products (Figure 1.20). Eq. 1.92 allows us to calculate ΔS at the temperature T from ΔS at a certain reference temperature T_0, and from ΔC_p. An analogous expression can be derived for isochoric processes, using C_V as reference value. For most biological systems that usually operate at constant pressure, these considerations are less relevant, however.

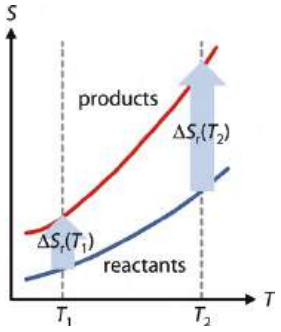


FIGURE 1.20: Temperature dependence of entropy and entropy changes. The entropies of the reactants and of the products depend on the temperature. In general, reactants and products will have different heat capacities, and thus a different dependence of their entropies on temperature. As a consequence, the change in entropy during the reaction, ΔS_r, the difference between $S_{products}$ and $S_{reactants}$, also depends on temperature (arrows).

1.5.4 The third law of thermodynamics and absolute entropy

We have discussed before that there is an absolute temperature minimum that cannot be reached (Section 1.3.3). At this absolute minimum in temperature, all matter would be perfectly ordered, which corresponds to an entropy of $S = 0$. Although $T = 0$ K cannot be reached, we can still formulate the limit of the entropy towards the absolute temperature minimum:

$$\lim_{T \to 0} S(T) = 0 \qquad \text{eq. 1.93}$$

This statement is the *third law of thermodynamics*. As a consequence of the third law, the *absolute entropy* of a compound as a function of temperature can be calculated by summing all entropy it has gained starting from $T = 0$ K (Figure 1.21; eq. 1.94). To do this, we have to integrate over the heat needed to increase the temperature of the compound in its solid state by one Kelvin, which is the heat capacity of the solid state $C_{p,s}$, integrated over the temperature range. At the melting temperature T_{melt}, heat will be absorbed during the phase transition to the liquid state without increase in temperature. Heat that is exchanged without changing the temperature of the system is termed *latent heat*. To calculate the associated entropy change, we have to divide the melting enthalpy ΔH_{melt} by the temperature T_{melt} according to the heat definition of the entropy. Similarly, at the boiling temperature T_{boil}, heat ΔH_{boil} will be absorbed during the transition to the gas phase, again without increasing the temperature until the phase transition is completed, accounting for an entropy change of $\Delta H_{boil}/T_{boil}$. In the liquid state, between T_{melt} and T_{boil}, the heat required for a temperature increase of one Kelvin is the heat capacity of the liquid state, $C_{p,l}$, above T_{boil} we have to use the heat capacity of the gaseous state, $C_{p,g}$, to integrate over the temperature range. Note that

the heat capacities for the different phases are different. By adding up all heat terms, divided by the temperature, we obtain the absolute entropy $S(T)$ as

$$S(T) = S(0) + \int_{0K}^{T_{melt}} \frac{C_{p,s}\,dT}{T} + \frac{\Delta H_{melt}}{T_{melt}} + \int_{T_{melt}}^{T_{boil}} \frac{C_{p,l}\,dT}{T} + \frac{\Delta H_{boil}}{T_{boil}} + \int_{T_{boil}}^{T} \frac{C_{p,g}\,dT}{T} \qquad \text{eq. 1.94}$$

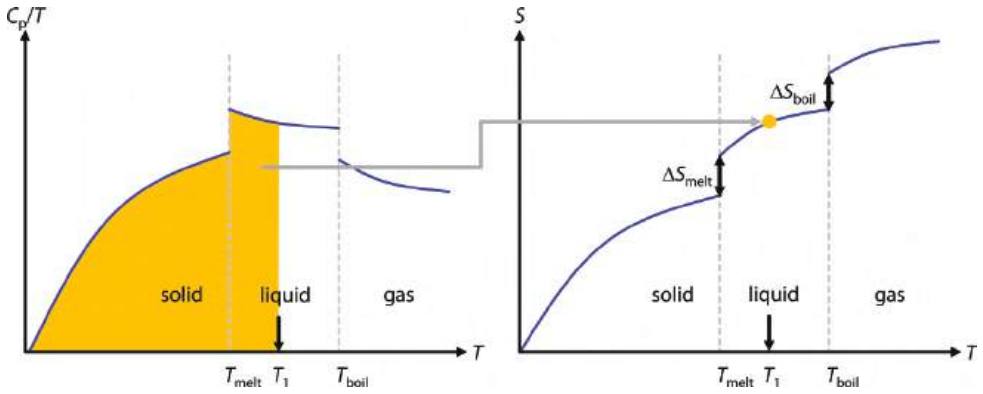


FIGURE 1.21: Absolute entropy. The absolute entropy of a compound can be calculated by summing all entropy it has gained from $T = 0$ K. Following the heat definition of entropy, the heat required to increase the temperature by one Kelvin, *i.e.* the heat capacity C_p ($C_{p,s}$, $C_{p,l}$, and $C_{p,g}$ for the solid, liquid, and gaseous states) has to be integrated over the relevant temperature ranges (left). Note that the heat capacities for the different phases are different. By integration over the temperature (orange area for T_1), the entropy as a function of temperature can be calculated (right; gray arrow, orange circle). At the melting temperature T_{melt}, heat will be absorbed during the phase transition to the liquid state without increase in temperature. The associated entropy increase ΔS_{melt} equals the melting enthalpy ΔH_{melt} divided by the temperature T_{melt}. Similarly, at the boiling temperature T_{boil}, heat ΔH_{boil} will be absorbed during the transition to the gas phase, again without increasing the temperature until the phase transition is completed, and the entropy increases by $\Delta S_{boil} = \Delta H_{boil}/T_{boil}$.

In biochemical processes, we are typically interested in changes in entropy, while the absolute entropy is of limited interest.

1.5.5 Entropy and order: the statistic interpretation

We introduced the state function entropy from the phenomenon of dissipation of matter and heat (Figure 1.19), and related entropy to order/disorder. In the following, we will see that the order or disorder of a particular state, and hence its entropy, can also be quantified by looking at the different ways by which a state can be generated. To illustrate this concept, we consider a system of four particles that can randomly distribute over two compartments (Figure 1.22). This system can exist in five different states: (1) all particles in the left compartment, (2) three particles left, one right, (3) two left, two right, (4) one left, three right, or (5) all four particles in the right compartment. The states (1) and (5), with all particles in one compartment, would be an "ordered" state, the state with two particles left, two right is what we usually perceive as random, or less ordered.

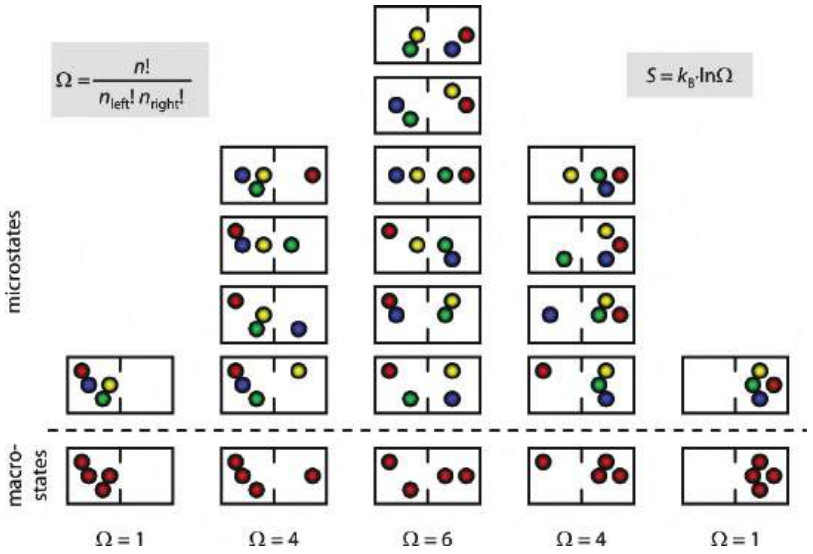


FIGURE 1.22: Macrostates, microstates, entropy, and statistics. Four particles (red, bottom) can be distributed over two compartments, left and right, such that all four particles are in the left compartment, three are left and one is in the right compartment, two each are left and right, one is left and three are right, or all four are in the right compartment. These states are called macrostates. In the case that we can distinguish the four particles (red, blue, green, yellow), there are different sub-states for each macrostate, the microstates. The macrostates with all four particles in the left or in the right compartment can only be generated in one defined way. The macrostates with the 3–1 distribution can be generated in four different ways, with either the red, the blue, the green, or the yellow particle separated from the others. The macrostate with two particles on each side can be generated in six different ways. The number of microstates depends on the total number of particles n and the number of particles in the left and right compartment, n_{left} and $n_{right} = n - n_{left}$. From the number of microstates, the entropy S can be calculated according to the Boltzmann equation (eq. 1.96) as $S = k_B \cdot \ln\Omega$, with the Boltzmann constant k_B.

How can order or randomness be expressed in quantitative terms? If we consider the same four-particle system, but now the four particles can be distinguished, *e.g.* by their different colors, we can tabulate the different arrangements to generate states (2) to (4) (Figure 1.22). For example, state (2) with one particle on the right can be generated in four different ways, with the red, yellow, green, or blue particle on the right-hand side. The same considerations lead to six possible ways to generate state (3), and again four different ways to generate state (4). There is still only one possibility to generate states (1) and (5). States 1–5 are the *macrostates* of the system. The different sub-states that lead to the net arrangement of particles in these states are called *microstates* (Figure 1.22; see Section 31.3). The microstates have the same energy, we say the macrostate is *degenerate* (see Section 3.2.2). We can calculate the number of microstates Ω from the total number of particles, n, and the number of particles in the left and right compartment, $n_{right} = j$ and $n_{left} = n - n_{right} = n - j$ as

$$\Omega = \frac{n!}{j!(n-j)!} \qquad \text{eq. 1.95}$$

The number of microstates is described by binomial coefficients that follow Pascal's triangle (coefficients for binomial distributions, see Section 31.3).

According to the statistical definition of entropy, the number of possible microstates (eq. 1.95) of each macrostate defines the entropy S of this state as

$$S = k_B \ln\Omega \qquad \text{eq. 1.96}$$

where k_B is the Boltzmann constant (Box 1.8). Eq. 1.96 directly relates the randomness of a state, or the number of possibilities by which it can be generated, to its entropy. In this respect, the entropy

is a measure of the order or randomness of a system: high entropy is related to a high degree of randomness or low order, and low entropy corresponds to a low degree of randomness or a highly ordered state.

We will revisit the concept of macro- and microstates in Chapter 3, when we discuss thermodynamics from a statistical point of view. Statistical thermodynamics can be used to describe structural changes of biomolecules (Section 3.3) as well as binding of ligands to biomolecules with multiple active sites (Sections 3.4).

BOX 1.8: THE BOLTZMANN CONSTANT k_B AND ITS RELATION TO THE GENERAL GAS CONSTANT.

In eq. 1.96, we have introduced the Boltzmann constant k_B as a proportionality constant between $\ln\Omega$ and the entropy S. In fact, k_B is directly related to the first constant we introduced, the general gas constant R, which is the proportionality constant between pV and nT in the ideal gas law (eq. 1.1). The relationship is very simple: k_B is the same constant per molecule that R is per mol. In other words:

$$R = N_A k_B \qquad \text{eq. 1.97}$$

where N_A is the Avogadro constant, the number of molecules per mol. By dividing R by N_A, we obtain $k_B = 8.3143\ \text{J mol}^{-1}\ \text{K}^{-1}/6.022{\cdot}10^{23}\ \text{mol}^{-1} = 1.381{\cdot}10^{-23}\ \text{J K}^{-1}$.

Are the definitions of entropy by heat and the statistical definition by order equivalent? We will apply both definitions to the reversible isothermal expansion of 1 mol of an ideal gas from V_1 to V_2 to confirm the equivalency. First, we consider the heat definition. The heat exchanged during the isothermal expansion is

$$\delta q = -\delta w = p\,\mathrm{d}V = \frac{RT}{V}\mathrm{d}V \qquad \text{eq. 1.98}$$

According to the heat definition, the entropy change ΔS associated with this process is thus

$$\Delta S = \frac{q^{\text{rev}}}{T} = \int \frac{\delta q^{\text{rev}}}{T} = \int_{V_1}^{V_2} \frac{R}{V}\mathrm{d}V = R\ln\frac{V_2}{V_1} \qquad \text{eq. 1.99}$$

Now we consider the statistical interpretation of the entropy. In the initial state, the gas occupies the volume V_1. We divide the volume V_1 into n_1 individual cells with a volume V (Figure 1.23). One particle then has n_1 cells it can occupy, giving n_1 possibilities to generate the macrostate with one particle in the volume V_1. If we consider two particles, the possibilities are $(n_1)^2$, and for n particles, they are $(n_1)^n$. In 1 mol of an ideal gas, we have N_A particles (where N_A is the Avogadro constant, $N_A = 6.022{\cdot}10^{23}\ \text{mol}^{-1}$) that can be arranged in $(n_1)^{N_A}$ different ways, giving rise to $(n_1)^{N_A}$ microstates for the initial state. For the final state, we can divide the volume V_2 into n_2 cells of the volume V. The same number of N_A particles can now be arranged in $(n_2)^{N_A}$ different ways; there are $(n_2)^{N_A}$ microstates for the final state. We can now relate the number of possibilities to generate the initial and final states to the entropy using the Boltzmann equation, and obtain

$$\Delta S = S_{\text{final}} - S_{\text{initial}} = k_B\ln\left(n_2^{N_A}\right) - k_B\ln\left(n_1^{N_A}\right) = k_B\ln\left(\frac{n_2}{n_1}\right)^{N_A} = N_A k_B\ln\left(\frac{n_2}{n_1}\right) \qquad \text{eq. 1.100}$$

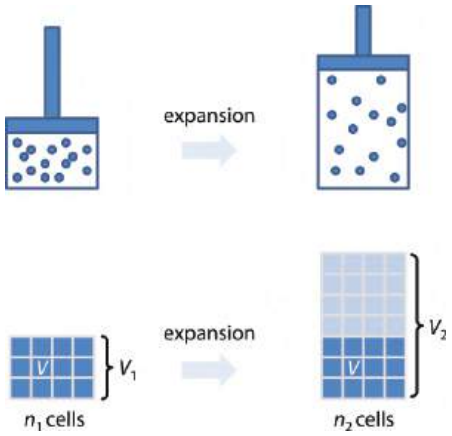


FIGURE 1.23: Statistical definition of entropy. We consider the reversible isothermal expansion of one mol of an ideal gas from volume V_1 (left) to volume V_2 (right). The volume V_1 can be divided into n_1 cells with a volume V, and the volume V_2 can be described as n_2 cells of the same volume V. The number of microstates for the initial state depends on the number of possibilities for N_A particles (N_A = Avogadro constant, $N_A = 6.022 \cdot 10^{23}$ mol^{-1}) to distribute themselves over the n_1 cells of volume V_1, which is $n_1^{N_A}$. The number of microstates for the final states is $n_2^{N_A}$. From the number of microstates, we can calculate the entropy of the initial and final states, and ΔS as the difference. ΔS calculated from the statistical interpretation of the entropy equals the entropy change calculated from the heat definition (see text).

From our definitions above, we can substitute n_2 and n_1 with V_2/V and V_1/V. Also, the product of N_A and k_B is the general gas constant R (Box 1.8). We can therefore rearrange eq. 1.100 to

$$\Delta S = R \ln\left(\frac{V_2 / V}{V_1 / V}\right) = R \ln \frac{V_2}{V_1} \qquad \text{eq. 1.101}$$

Thus, we obtain the same result for the entropy change ΔS using the statistical (eq. 1.101) or the heat definition (eq. 1.99) of the entropy S, which illustrates that both concepts are equivalent.

1.6 THE FREE ENERGY *G*: COMBINING SYSTEM AND SURROUNDINGS

In Section 1.5.2, we derived the Clausius inequality (eq. 1.83), which tells us that spontaneous processes are associated with an increase in entropy and that the entropy of an isolated system can only increase. By definition, isolated systems do not exchange heat with the surroundings, and entropy considerations can be limited to the system itself. In contrast, a closed system exchanges heat with its surroundings. To describe entropy changes associated with processes in closed systems, we therefore have to include the surroundings in our entropy considerations: we have to take into account the entropy change of the system itself, ΔS_{system}, and of the surroundings, $\Delta S_{surroundings}$. The total change in entropy ΔS_{total} then becomes

$$\Delta S_{total} = \Delta S_{system} + \Delta S_{surroundings} \qquad \text{eq. 1.102}$$

For a reversible process, an entropy reduction within the system will be compensated by an increase in entropy of the surroundings of the same magnitude:

$$\Delta S_{surroundings} = -\Delta S_{system} \qquad \text{eq. 1.103}$$

This change in entropy is also termed *entropy transport* (Box 1.7). For an irreversible process, however, more entropy will be generated in the surroundings (*entropy production*; Box 1.7).

$$\Delta S_{\text{surroundings}} > -\Delta S_{\text{system}} \qquad \text{eq. 1.104}$$

Instead of calculating entropy changes of the system and the surroundings separately, it would be convenient to combine these individual entropy changes for the system and the surroundings in one state function. Starting from the total change in entropy (eq. 1.102), we can express the change in entropy of the surroundings as the ratio of the heat transferred to or from the system, and the temperature at which this heat transfer takes place:

$$\Delta S_{\text{total}} = \Delta S_{\text{system}} - \frac{q_{\text{system}}}{T} \qquad \text{eq. 1.105}$$

For p = const., the heat exchanged in this process equals ΔH. At constant temperature T, we can therefore write

$$\Delta S_{\text{total}} = \Delta S_{\text{system}} - \frac{\Delta H_{\text{system}}}{T} \qquad \text{eq. 1.106}$$

The negative sign for q and ΔH results from the system-centered perspective: the heat taken up by the surroundings is released from the system ($\delta q < 0$), and heat absorbed from the surroundings is taken up by the system ($\delta q > 0$).

By rearranging eq. 1.106 we obtain

$$-T\Delta S_{\text{total}} = \Delta H_{\text{system}} - T\Delta S_{\text{system}} \qquad \text{eq. 1.107}$$

On the right-hand side of eq. 1.107 we now only find changes in thermodynamic parameters of the system. Why is this relation useful? A closed system and its surroundings together constitute an isolated system, and ΔS_{total} represents the change in entropy in this isolated system. For spontaneous processes in isolated systems, ΔS is always positive, meaning ΔS_{total} must be positive for any spontaneous process that occurs in a closed system. The temperature T, the absolute temperature in K, is always positive. For a positive ΔS_{total}, the term $-T\Delta S_{\text{total}}$ on the left-hand side in eq. 1.107 becomes negative. We name $-T\Delta S_{\text{total}}$ the change in free energy ΔG associated with the spontaneous process. ΔG summarizes entropy changes in system and surroundings. Our criterion for spontaneous processes, an increase in total entropy S_{total} or a negative $-T\Delta S_{\text{total}}$, can therefore now be expressed in terms of ΔG: spontaneous processes are characterized by a decrease in the free energy ($\Delta G < 0$). Processes in equilibrium do not involve a change in free energy ($\Delta G = 0$), and processes associated with an increase in free energy ($\Delta G > 0$) do not occur spontaneously. Processes with $\Delta G < 0$ are termed *exergonic*, processes with $\Delta G > 0$ are *endergonic*.

With ΔG, we have implicitly introduced a new state function, the *free energy* G. The general definition of the free energy is

$$G = H - TS \qquad \text{eq. 1.108}$$

G is also called the *Gibbs free energy*. Complete differentiation of G gives

$$\mathrm{d}G = \mathrm{d}H - T\,\mathrm{d}S - S\,\mathrm{d}T \qquad \text{eq. 1.109}$$

For isothermal processes ($\mathrm{d}T = 0$), eq. 1.109 simplifies to

$$\mathrm{d}G = \mathrm{d}H - T\,\mathrm{d}S \qquad \text{eq. 1.110}$$

For measurable changes of ΔG during a reaction, we can formulate the analogous relationship

$$\Delta G = \Delta H - T\Delta S \qquad \text{eq. 1.111}$$

Eq. 1.111 is called the *Gibbs-Helmholtz equation*. Comparison of eq. 1.111 and eq. 1.107 shows that the right-hand sides are identical, and refer to the enthalpic and entropic changes within the closed system.

Changes in state functions can be calculated from the difference between the values for the final and initial states. For a chemical reaction, the changes in free energy ΔG_r therefore correspond to the difference in the free energies of the products (final state) and the reactants (initial state):

$$\Delta G_r = G_{products} - G_{reactants} \qquad \text{eq. 1.112}$$

To calculate the values of the respective state function for reactants and products, we can use the elements as a reference state. This way, we can define the value of the state function for a molecule as the change in this state function when the molecule is generated from the elements (Figure 1.24). For example, the free energy G of a compound can be defined relative to the free energy of the elements from which this compound can be generated. This value is called the *standard free energy of formation*, ΔG_f. From the standard free energy of formation, the change in free energy associated with a chemical reaction can be determined as

$$\Delta G_r = \sum G_{f,products} - \sum G_{f,reactants} \qquad \text{eq. 1.113}$$

Analogous to the standard free energy of formation ΔG_f, standard enthalpies and entropies of formation, ΔH_f and ΔS_f, can be defined, and their values are tabulated in reference books.

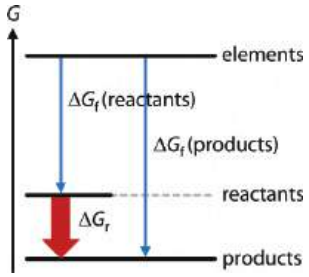


FIGURE 1.24: The standard free energy of formation and the change in free energy during a reaction. The free energy *G* is a state function, and its change during a chemical reaction can be calculated from the difference of the free energies of products and reactants (red arrow). The free energies of products and reactants can be defined relative to the elements. The free energy of a compound relative to its constituent elements is the free energy of formation, ΔG_f (blue arrows).

It is important to note the relationship between forward and reverse reactions in terms of changes in state functions: if a reaction is associated with a certain $\Delta G_{forward}$, the reverse reaction will be accompanied by $\Delta G_{reverse}$, the opposite change in ΔG:

$$\Delta G_{forward} = -\Delta G_{reverse} \qquad \text{eq. 1.114}$$

Hence, if a forward reaction occurs spontaneously ($\Delta G_{forward} < 0$), the reverse reaction will not be spontaneous ($\Delta G_{reverse} > 0$), and *vice versa*.

1.6.1 Entropy- and enthalpy-driven reactions

We have now derived a decrease in ΔG as the criterion for a spontaneous reaction. According to the Gibbs-Helmholtz equation (eq. 1.111), ΔG is calculated as the difference between the change in enthalpy, ΔH, and the change in entropy, ΔS, converted into an energy value by multiplication with T (eq. 1.111). It is the balance of the ΔH and the $-T\Delta S$ terms that determines whether a reaction is spontaneous. A large negative change in enthalpy and a large increase in entropy are both favorable

and lead to a negative value for ΔG. Likewise, an increase in enthalpy and a decrease in entropy are both unfavorable and ΔG becomes positive. However, ΔH and ΔS can also have opposite effects. If the unfavorable term (positive ΔH or negative ΔS = positive $-T\Delta S$) overcompensates the favorable term (positive ΔS = negative $-T\Delta S$ or negative ΔH), ΔG is positive, and the reaction does not occur spontaneously. If, however, the favorable term prevails and overcompensates the unfavorable term, ΔG becomes negative, and the reaction will occur spontaneously. Reactions that are energetically favored because of a large negative enthalpy change, but associated with a decrease in entropy, are referred to as *enthalpy-driven reactions*. Many biological association processes, from simple, binary protein-ligand interactions to the formation of multi-component complexes such as ribosomes, fall into this category. In contrast, spontaneous reactions associated with an increase in enthalpy that are favored by a large increase in entropy are called *entropy-driven reactions*. The dissolution of $NaNO_3$ (Section 1.5.1) is an example for such an endothermic, yet exergonic and, hence, spontaneous reaction. As the entropy change in the Gibbs-Helmholtz equation (eq. 1.111) is weighted by temperature, an unfavorable reaction may become entropy-driven at a higher temperature, or enthalpy-driven at a lower temperature (Figure 1.25). For example, proteins are thermodynamically stable (ΔG for unfolding > 0) only in a certain temperature range, typically around the physiological temperature of the respective organism. At high temperatures, ΔG for unfolding becomes negative because the positive entropic term ($T\Delta S$) becomes large enough to overcompensate the unfavorable enthalpic term; the protein unfolds in an entropy-driven process (heat denaturation). At low temperatures, on the other hand, ΔG for unfolding becomes negative because the negative enthalpic term (ΔH) overcompensates the small negative entropic term ($T\Delta S$); the protein unfolds in an enthalpy-driven process (Box 1.9, see Figure 1.26).

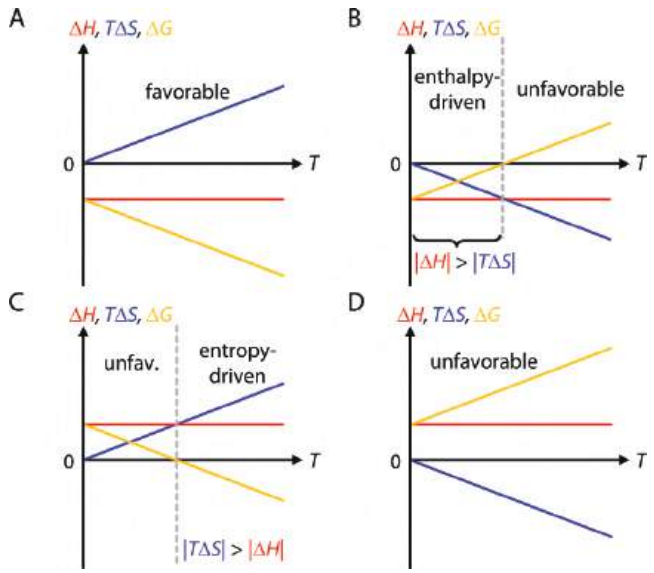


FIGURE 1.25: Temperature dependence of enthalpy and entropy changes and effects on the spontaneity of the reaction. A: $\Delta H < 0$, $\Delta S > 0$. The reaction is favorable ($\Delta G < 0$) at all temperatures. B: $\Delta H < 0$, $\Delta S < 0$. The reaction is unfavorable at higher temperatures, but becomes favorable at low temperatures (enthalpy-driven). C: $\Delta H > 0$, $\Delta S > 0$. The reaction is unfavorable at low temperatures, but becomes favorable (entropy-driven) at higher temperatures. D: $\Delta H > 0$, $\Delta S < 0$. The reaction is unfavorable ($\Delta G > 0$) at all temperatures.

The free energy change ΔG expresses the balance between enthalpic and entropic changes. Many processes are associated with enthalpic (ΔH) and entropic terms ($-T\Delta S$) that are substantial, but of opposite sign, and the resulting ΔG can be very modest. This effect is called *enthalpy-entropy compensation*. It is often encountered in biochemical processes, including protein folding (Box 1.9), hybridization of nucleic acids, protein-ligand and protein-protein interactions (see Sections 18.3.6.2 and 18.3.6.3), or the association of lipids into monolayers, micelles, and membranes (see Section 18.4.3).

BOX 1.9: ENTROPY-ENTHALPY COMPENSATION IN PROTEIN FOLDING.

The three-dimensional structure of proteins is stabilized by numerous individual interactions involving side-chain and backbone atoms (Section 18.3), which constitutes a favorable (negative) enthalpic contribution to protein folding ($\Delta H \ll 0$). On the other hand, protein folding converts a flexible, disordered polypeptide chain into a highly ordered three-dimensionally structured protein; the entropic contribution for folding is unfavorable (negative; $\Delta S \ll 0$). Entropic and enthalpic changes upon folding are substantial (hundreds of kJ mol^{-1}), but the resulting thermodynamic stability ΔG, the difference between ΔH and $T\Delta S$ (eq. 1.111), is often very moderate (a few to tens of kJ mol^{-1}) because of *enthalpy-entropy compensation* (see Figure 1.26). We will see in the following Section 1.6.2 that the temperature dependence of ΔH and ΔS (and thus of ΔG) leads to enthalpy-driven protein unfolding at low temperatures, and entropy-driven unfolding at high temperatures (Figure 1.26).

1.6.2 Pressure and temperature dependence of the free energy

We will now derive the pressure- and temperature dependence of the free energy *G*. Starting from the definition of the free energy (eq. 1.108), we can differentiate *G* to

$$\mathrm{d}G = \mathrm{d}H - T\mathrm{d}S - S\mathrm{d}T \qquad \text{eq. 1.115}$$

From the definition of the enthalpy, dH can be calculated as

$$\mathrm{d}H = \mathrm{d}U + p\mathrm{d}V + V\mathrm{d}p \qquad \text{eq. 1.59}$$

and by substitution we obtain

$$\mathrm{d}G = \mathrm{d}U + p\mathrm{d}V + V\mathrm{d}p - T\mathrm{d}S - S\mathrm{d}T \qquad \text{eq. 1.116}$$

We can further substitute dU by

$$\mathrm{d}U = \delta q + \delta w \qquad \text{eq. 1.4}$$

giving

$$\mathrm{d}G = \delta q + \delta w + p\mathrm{d}V + V\mathrm{d}p - T\mathrm{d}S - S\mathrm{d}T \qquad \text{eq. 1.117}$$

Now we include the (heat) definition of the entropy (eq. 1.74), and substitute TdS for δq. Note that this restricts our considerations to reversible processes, because otherwise the equality $T\mathrm{d}S = \delta q$ does not hold. We further restrict our considerations to expansion work, and can substitute $-p$dV for δw. Altogether, we obtain

$$\mathrm{d}G = T\mathrm{d}S - p\mathrm{d}V + p\mathrm{d}V + V\mathrm{d}p - T\mathrm{d}S - S\mathrm{d}T \qquad \text{eq. 1.118}$$

The TdS and pdV terms in eq. 1.118 cancel and we arrive at

$$\mathrm{d}G = V\mathrm{d}p - S\mathrm{d}T \qquad \text{eq. 1.119}$$

For T = const. (dT = 0), eq. 1.119 further reduces to

$$dG = V dp \qquad \text{eq. 1.120}$$

From eq. 1.120, we can derive the pressure dependence of G by integration from an initial pressure p^0 to the actual pressure p:

$$G = \int_{G(p^0)}^{G(p)} dG = \int_{p^0}^{p} V\, dp \qquad \text{eq. 1.121}$$

Substitution of V according to the ideal gas equation (eq. 1.1; $p = nRT/V$) and evaluation of the integrals gives the pressure dependence of G as

$$G(p) = G(p^0) + nRT \ln \frac{p}{p^0} \qquad \text{eq. 1.122}$$

Eq. 1.122 allows us to calculate G for any given pressure p if we know G at the standard pressure p^0. By analogy, we can express the pressure dependence of changes in the free energy, ΔG, as

$$\Delta G(p) = \Delta G(p^0) + nRT \ln \frac{p}{p^0} \qquad \text{eq. 1.123}$$

The temperature dependence of G and ΔG can, in principle, also be derived from eq. 1.119. At constant pressure ($dp = 0$), eq. 1.119 simplifies to

$$dG = -S\, dT \qquad \text{eq. 1.124}$$

Eq. 1.124 is the derivative of the Gibbs-Helmholtz equation (eq. 1.111), with the assumption that H is independent of temperature. If S were also independent of temperature, we could directly integrate eq. 1.124 to obtain $G(T)$. However, we have seen in Sections 1.4 and 1.5.3 that both entropy and enthalpy are temperature-dependent. The temperature dependence of G is therefore calculated directly from $H(T)$ and $S(T)$ at any given temperature T using the Gibbs-Helmholtz-equation (eq. 1.111):

$$G(T) = H(T) - T \cdot S(T) \qquad \text{eq. 1.125}$$

We have related the temperature dependence of H and S at constant pressure to the heat capacity C_p in eq. 1.64 and eq. 1.91. By substituting these equations into eq. 1.125, we obtain

$$G(T) = H(T_0) + C_p(T - T_0) - T\left(S(T_0) + C_p \ln \frac{T}{T_0}\right) \qquad \text{eq. 1.126}$$

Rearranging eq. 1.126 and summarizing $H(T_0)$–$T{\cdot}S(T_0)$ as $G(T_0)$ gives

$$G(T) = G(T_0) + C_p\left(T - T_0 - T \ln \frac{T}{T_0}\right) \qquad \text{eq. 1.127}$$

Again, we can write the analogous expression for the change of G during a reaction, $\Delta G(T)$:

$$\Delta G(T) = \Delta G(T_0) + \Delta C_p\left(T - T_0 - T \ln \frac{T}{T_0}\right) \qquad \text{eq. 1.128}$$

Thus, if ΔH, ΔS, or ΔG are known at a reference temperature T_0, their values $\Delta S(T)$, $\Delta H(T)$, and $\Delta G(T)$ at the temperature T can be calculated using the heat capacity ΔC_p. In other words: the heat capacity ΔC_p determines the temperature dependence of ΔG, ΔH, and ΔS. We will revisit the

temperature and pressure dependence of G and ΔG later when we discuss the effect of temperature and pressure on equilibrium constants (Section 2.1).

Eq. 1.128 gives a bell-shaped profile for $\Delta G(T)$ as a function of temperature (Figure 1.26). The curvature of this curve is determined by ΔC_p. ΔC_p thus determines the range where ΔG is negative, and where it becomes positive. A prominent example of the importance of changes in heat capacity is protein folding. ΔC_p determines the temperature range in which ΔG of unfolding is positive, and thus defines the temperature range where the protein is stably folded. The curve of ΔG_{unfold} as a function of temperature is also called the protein stability curve (Figure 1.26; Section 18.5.4). As a consequence of the curvature of the stability curve, proteins do not only unfold at high temperatures (heat denaturation) but also at low temperatures (cold denaturation). Heat denaturation is an entropy-driven reaction, cold denaturation is enthalpy-driven (Section 1.6.1; Box 1.9).

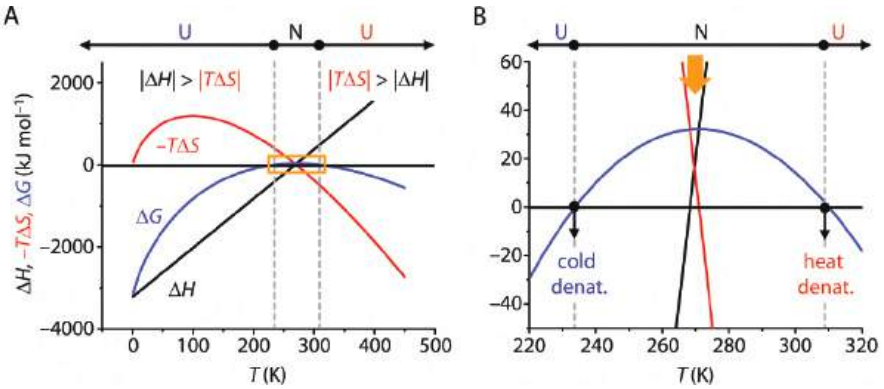


FIGURE 1.26: Stability curve of a protein. A: Temperature dependence of ΔH (black), $-T\Delta S$ (red), and ΔG (blue) for unfolding of a protein. $\Delta H(T_m)$ = 500 kJ mol^{-1}, ΔC_p = 12 kJ mol^{-1} K^{-1}, T_m = 310 K. At low temperatures, ΔG for unfolding is negative because of a large negative ΔH term, the protein unfolds in an enthalpy-driven reaction (U, unfolded). At high temperatures, ΔG is negative because of a large negative $-T\Delta S$ term; the protein unfolds in an entropy-driven reaction (U). In the temperature range in-between, ΔG is positive, and the folded protein (N, native) is stable. The strong enthalpy-entropy compensation leads to moderate values for ΔG in the physiologically relevant temperature range (orange rectangle, magnified in B). B: Physiologically relevant range from panel A. At the maximum of the bell-shaped ΔG curve, the thermodynamic stability of the folded protein is maximal (orange arrow). As a consequence of the curvature, ΔG of unfolding becomes negative at high and low temperatures, *i.e.* unfolding becomes a spontaneous reaction, and the unfolded state is the more stable state at high and at low temperatures (heat and cold denaturation).

Heat denaturation of proteins has its molecular origin (see Section 1.6.4) in the temperature dependence of the interactions stabilizing the folded state: these stabilizing interactions are dominated by a favorable enthalpic term, and hence become weaker (*i.e.* less stabilizing) with increasing temperature, until the favorable entropic term for unfolding dominates and protein unfolds in an entropy-driven reaction. There is one exception, however, which is the hydrophobic effect. Its stabilizing effect is based on the exclusion of hydrophobic side chains from the polar, aqueous solvent (Section 18.5.1). The hydrophobic effect is an entropic effect. It becomes less stabilizing with decreasing temperatures, until finally the solvation of hydrophobic side chains becomes thermodynamically favorable, and constitutes the driving force for enthalpy-driven cold denaturation.

1.6.3 Standard states

We have seen in the previous Sections that we can calculate the values of different state functions relative to certain reference states. We can calculate H, S, and G at a temperature T relative to the reference temperature T_0 (eq. 1.64, eq. 1.91, and eq. 1.127), and we can calculate G at a pressure p relative to the standard pressure p^0 (eq. 1.122). It therefore makes sense to define common reference states, which are called standard states. The pressure under standard conditions, p^0, is 10^5 Pa (1 bar),

and the standard temperature T_0 is 298.15 K (25°C). Values of the state functions and their changes at standard conditions are marked by a superscript "0", as in ΔH^0, ΔS^0, ΔG^0. We can now use the equations we derived for the temperature and pressure dependence of the state functions, and calculate their values at any pressure p or temperature T, relative to the standard state.

We will see in Sections 1.7.5 and 1.7.6 that we can also calculate the concentration dependence of state functions relative to a standard state, which is required to describe the thermodynamics of mixtures. The standard state for components of mixtures is the pure component: pure solid, pure liquid, or pure gas, depending on the aggregation state of the pure compound at T_0 and p^0. For solutions, the standard state of the solvent is the pure solvent. In contrast, the standard state for the dissolved components, the *solutes*, is a concentration of 1 M.

The definition of the standard states for solutes has certain limitations when we look at the thermodynamics of biochemical reactions. Biochemical reactions at physiological conditions usually occur at a neutral pH value of around 7. Yet, the definition of the standard state refers to 1 M concentration of the solute. For H^+ ions, this corresponds to a solution with pH 0, a condition that is of little physiological relevance. Therefore, the biochemical standard state uses a different definition for protons: the standard state is pH 7 (10^{-7} M protons), whereas for all other solutes, the standard state of 1 M concentration is maintained. The corresponding values for the state functions under biochemical standard conditions are denoted by a prime, and written as $\Delta H^{0\prime}$, $\Delta S^{0\prime}$, and $\Delta G^{0\prime}$. It is important to note that it does not matter which standard state we choose to calculate the value of the state function at a certain concentration, as long as we consistently stay in one of the reference systems (see Section 1.7.5, Box 1.12).

1.6.4 Relation of free energy, enthalpy, and entropy to molecular properties

How do molecular properties contribute to enthalpy and entropy, and how can changes in these state functions during chemical reactions be related to molecular properties of the compounds involved? Enthalpy is the sum of all energies of each molecule in the system (intramolecular energy) and of the energy from interactions between these molecules (intermolecular energy; Figure 1.27). The intramolecular energy contains contributions from translational, rotational, and vibrational motions of each individual molecule (also referred to as the thermal energy), and the energy is stored in its chemical bonds. In macromolecules, intramolecular interactions also contribute to the intramolecular energy. The intermolecular energy contains contributions from van der Waals- and ionic interactions and intermolecular hydrogen bonds (Section 17.2.2). Entropy, which we have introduced as randomness or ways of realization of a certain state in Section 1.5.5, cannot be defined on the level of an individual molecule. Instead, entropy is a measure of the different ways to distribute the energy of the entire system among its individual molecules. Energies associated with translation, rotation, and vibration can be distributed over each molecule (vibration) or among molecules (translation, rotation). In contrast, the energy of chemical bonds is fixed within each molecule and cannot be distributed within or among molecules. Bonds therefore do not contribute to entropy. Box 1.10 illustrates how changes in entropy and enthalpy during a chemical reaction can be related to molecular properties. The connection between molecular properties and thermodynamic properties of a system is at the heart of statistical thermodynamics (Chapter 3).

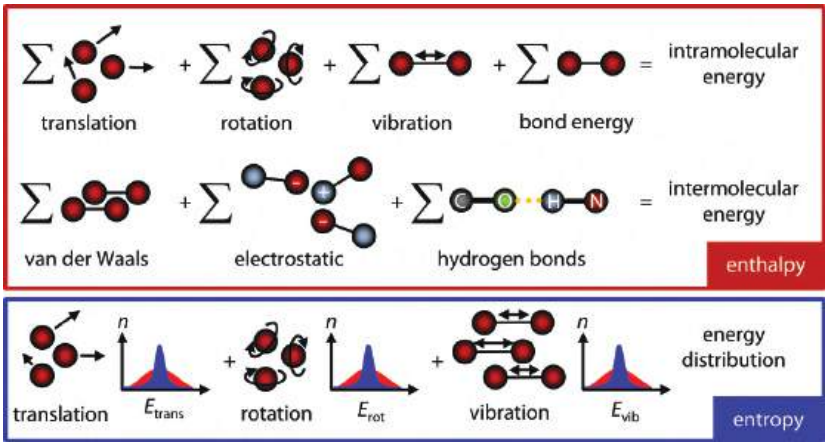


FIGURE 1.27: Molecular components of entropy and enthalpy. Enthalpy is the sum of all intramolecular and intermolecular energies, such as translation, rotation, and vibration of individual molecules and the energy of covalent bonds, or van der Waals- and hydrogen bonding and ionic interactions between molecules. Entropy can only be defined on the basis of an ensemble of molecules and is related to the distribution of energy among molecules. While translational, rotational, and vibrational energies can be distributed among molecules, and contribute to the entropy of the system, energy in chemical bonds is located and cannot be distributed, and therefore does not contribute to entropy.

BOX 1.10: MOLECULAR CONTRIBUTIONS TO ENTHALPY AND ENTROPY CHANGES IN ATP HYDROLYSIS.

ATP is hydrolyzed to ADP and inorganic phosphate. The reaction is associated with a free energy change at pH 7 of $\Delta G^{0'}$ of –31 kJ mol^{-1} (310 K). The large negative $\Delta G^{0'}$ results from a negative enthalpy change ($\Delta H^{0'}$ = –20.1 kJ mol^{-1}) and a large increase in entropy ($\Delta S^{0'}$ = +33.5 J mol^{-1} K^{-1}), *i.e.* both enthalpy and entropy changes are favorable. What are the molecular reasons for these changes? We first take a look at the enthalpy. The magnitude of the enthalpy change is dominated by the disruption and formation of bonds. During ATP hydrolysis, one P-O bond is broken (between γ- and β-phosphate; Figure 1.28), which is associated with a positive $\Delta H^{0'}$. However, a new P-O bond is formed between the water molecule that performs the nucleophilic attack and the γ-phosphate, which is associated with a negative $\Delta H^{0'}$. The energies involved in bond breaking and bond formation are large, but because they are equal in magnitude but of opposite sign, they cancel and do not contribute to the overall $\Delta H^{0'}$. Apart from bonds being broken and formed, ATP hydrolysis leads to a distribution of the negative charges of the three phosphate groups (ATP^{4-} at pH 7) onto two separate molecules, ADP (ADP^{2-}) and phosphate (HPO_4^{2-}). The separation reduces the electrostatic repulsion, which is energetically favorable and associated with a negative $\Delta H^{0'}$. Finally, we have to consider hydration because the reaction occurs in aqueous solution. The hydration of ADP and phosphate is energetically more favorable compared to the hydration of ATP, and hydration also contributes to the negative $\Delta H^{0'}$. The negative change in enthalpy during ATP hydrolysis is thus a combined effect of the reduced electrostatic repulsion and the increased hydration of the products.

The entropic term depends on the possibilities to distribute energy among the molecules. ATP hydrolysis converts one reactant, ATP, into two products. The overall energy can now be distributed over both molecules, which causes an increase in entropy. On the other hand, the two product molecules are hydrated. Although hydration contributes favorably to the enthalpy change, it also leads to a higher order of water molecules, associated with a decrease

(*Continued*)

in water entropy. The overall increase in entropy for the hydrolysis reaction tells us that the favorable entropy term, because of splitting one molecule into two, overcompensates the unfavorable term due to water ordering. The decrease in enthalpy and increase in entropy then add up to an overall decrease in free energy, and ATP hydrolysis is a spontaneous reaction (under standard conditions).

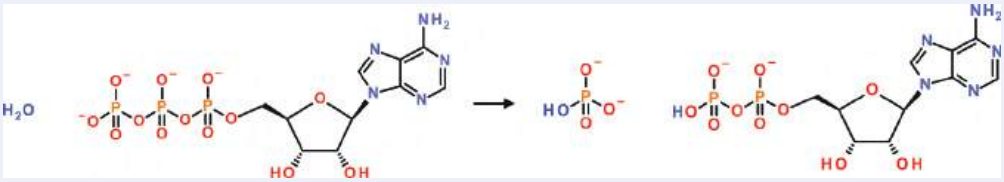


FIGURE 1.28: Hydrolysis of ATP to ADP and phosphate. ATP hydrolysis to ADP and phosphate is associated with a decrease in enthalpy ($\Delta H^{0\prime} = -20.1$ kJ mol^{-1}), and an increase in entropy ($\Delta S^{0\prime} = 33.5$ J mol^{-1} K^{-1}), leading to an overall decrease in free energy ($\Delta G^{0\prime} = -31$ kJ mol^{-1}) at 310 K.

Note that the term "high-energy bond" often used to describe the large change in free energy associated with ATP hydrolysis is very misleading. On a molecular level, bond breaking and formation is isoenergetic, and does not contribute to the overall $\Delta H^{0\prime}$ and $\Delta G^{0\prime}$. It is also important to note that the large, negative $\Delta G^{0\prime}$ does not tell us anything about the velocity of ATP hydrolysis. In fact, ATP hydrolysis is a slow reaction because of a high activation barrier (Section 15.2). In the cell, hydrolysis of ATP is faster because it is catalyzed by enzymes. Synthesis and hydrolysis of ATP, and thus the energy level of the cell, can be regulated thereby.

1.7 THE CHEMICAL POTENTIAL μ

So far, we have considered processes in isolated or closed systems that do not exchange matter with their surroundings. In other words, in the processes we discussed, the amount of substance n (in mol) has remained constant. However, we often need to treat processes where n is changing, *e.g.* chemical reactions or open systems. Open systems exchange matter with their surroundings, and their composition is not constant. Chemical reactions convert reactants into products. Over the course of the reaction, the reactants are consumed, and their amount n is reduced. Conversely, products are formed, and their amount n keeps increasing as long as the reaction proceeds. We have seen that U, H, S, and G are extensive state functions: their values depend on the size of the system, and thus on n. The free energy G is therefore not necessarily a useful state function when we consider chemical reactions and open systems where n is not constant.

We can introduce a new state function that is independent of the amount n by the following consideration: if we consider a pure compound, an infinitesimal change in the amount n of this compound by dn leads to an infinitesimal change dG in the free energy of the system (at constant pressure and temperature). We define the proportionality constant as the *chemical potential* μ of this compound:

$$\mathrm{d}G = \mu\, \mathrm{d}n \qquad \text{eq. 1.129}$$

or

$$\mu = \left(\frac{\partial G}{\partial n} \right)_{p,T} \qquad \text{eq. 1.130}$$

Eq. 1.130 is a partial derivative (see Section 31.6.2): G is a function of p, T, and n, but with the partial derivative we only consider changes in n. For a pure compound, the chemical potential is simply equal to the free energy G per mol, or molar free energy G_m:

$$\mu = \frac{G}{n} = G_m \qquad \text{eq. 1.131}$$

In analogy to eq. 1.131, we can also define the molar enthalpy H_m, the molar entropy S_m, or the molar volume V_m of a pure substance by dividing the corresponding extensive properties by the amount of substance n. By the normalization to one mol, the extensive properties G, H, S, and V become intrinsic properties G_m, H_m, S_m, or V_m.

1.7.1 The chemical potential as a driving force for chemical reactions

The general considerations on the chemical potential apply to pure substances, but it becomes much more useful for the description of mixtures. Most chemical reactions occur in mixtures of reactants and generate a mixture of products. ΔH_r, ΔS_r, and ΔG_r describe the changes in thermodynamic state functions for the reaction, but summarize the entire system. To understand the behavior of mixtures, it is often useful to describe the state from the perspective of one individual component. The chemical potential provides a convenient way to quantify the contribution of each component of the mixture to the free energy. For a mixture of i components, we can sum the changes in free energy dG of the entire mixture caused by the increase in the amount n_i of each component as

$$dG = \sum_i \mu_i \, dn_i \qquad \text{eq. 1.132}$$

where μ_i is the chemical potential of compound i. From eq. 1.132, we can infer that the free energy G of a mixture can be calculated from the chemical potential of the individual components i of the system and their amount n_i (eq. 1.133):

$$G = \sum_i \mu_i n_i \qquad \text{eq. 1.133}$$

Changes in free energy during a reaction are defined as the difference between G of the final state (products) and G of the initial state (reactants), leading to

$$\Delta G = G_{products} - G_{reactants} = \sum_{products} \mu_i n_i - \sum_{reactants} \mu_i n_i \qquad \text{eq. 1.134}$$

or

$$\Delta G = \sum_i \mu_i \Delta n_i \qquad \text{eq. 1.135}$$

Δn_i denotes changes in the amount of substance for each component, with a positive sign for products that are being formed, and a negative sign for reactants being consumed. This convention is again a reflection of the system-centered perspective: reactants are disappearing from the system, and products are appearing. The value for ΔG in eq. 1.134 becomes negative if the sum of the chemical potential of the reactants is larger than the chemical potential of the products. For open systems and for chemical reactions, we can therefore formulate the condition $\Delta G < 0$ for spontaneous processes in terms of the chemical potentials of the reactants and the products as

$$\sum_{\text{reactants}} \mu_i n_i > \sum_{\text{products}} \mu_i n_i \qquad \text{eq. 1.136}$$

A reaction with

$$\sum_{\text{reactants}} \mu_i n_i = \sum_{\text{products}} \mu_i n_i \qquad \text{eq. 1.137}$$

is in equilibrium, and a reaction with

$$\sum_{\text{reactants}} \mu_i n_i < \sum_{\text{products}} \mu_i n_i \qquad \text{eq. 1.138}$$

does not occur spontaneously (Table 1.2). Eq. 1.136 explains the term "chemical potential" for μ. A potential is generally defined as the capability to perform work. In accordance with this definition, the chemical potential is a measure of the capability to drive a chemical reaction from reactants to products (Figure 1.29).

TABLE 1.2:
Criteria for spontaneous processes, equilibria, and non-spontaneous processes in isolated, closed, and open systems.

	spontaneous	equilibrium	non-spontaneous
isolated	$\Delta S > 0$	$\Delta S = 0$	$\Delta S < 0$
closed	$\Delta G < 0$	$\Delta G = 0$	$\Delta G > 0$
open	$\sum n\mu_{\text{reactants}} > \sum n\mu_{\text{products}}$	$\sum n\mu_{\text{reactants}} = \sum n\mu_{\text{products}}$	$\sum n\mu_{\text{reactants}} < \sum n\mu_{\text{products}}$

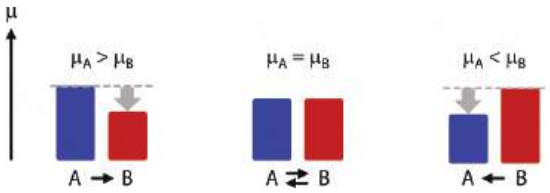


FIGURE 1.29: The chemical potential as a driving force for reactions. The difference in chemical potential between reactant A (blue) and product B (red) is the driving force for a reaction from A to B (left, gray arrow). When the chemical potentials of A and B are identical (middle), there is no driving force for the reaction, and the reaction is in equilibrium. A higher chemical potential of B is the driving force for the reverse reaction from B to A (right, gray arrow).

1.7.2 The chemical potential and stable states: phase diagrams

The chemical potential is a driving force for processes and chemical reactions. We have derived the chemical potential from the free energy, and have seen that the chemical potential of a pure substance is equivalent to the molar free energy. For pure substances, we can therefore describe the chemical potential according to the Gibbs-Helmholtz-equation (eq. 1.111), using the molar enthalpy H_m and the molar entropy S_m as parameters:

$$\mu = G_m = H_m - TS_m \qquad \text{eq. 1.139}$$

We now assume that H_m and S_m do not depend on temperature. Although this is a simplification (we have learnt that the heat capacity C_p describes the temperature dependence of both H and S; eq. 1.64 and eq. 1.91), it helps us to understand why compounds exist in solid, liquid, or gaseous states, why these states are stable in certain temperature ranges, and why they interconvert at defined

temperatures. For temperature-independent H_m and S_m, eq. 1.139 describes a linear dependence of μ on the temperature, with a slope $-S_m$ and a y-axis intercept H_m. Thus, when we plot μ as a function of T (Figure 1.30), the chemical potentials for the solid, liquid, and gaseous states of a pure substance are straight lines with different slopes and intercepts. To assign differences in slopes and intercepts, we have to compare the molar enthalpies and entropies for the three states. The solid phase has the highest order, and the order decreases from solid (s) to liquid (l) to gaseous state (g), which is equivalent to an increase in randomness from solid to gas. Therefore we can write for the molar entropies

$$S_m(g) > S_m(l) > S_m(s) \qquad \text{eq. 1.140}$$

Melting a solid requires energy to overcome interactions between particles in the solid phase. Likewise, evaporation of a liquid to a gas requires energy to overcome interactions between particles in the liquid. This energy remains in the system, which means that

$$H_m(g) > H_m(l) > H_m(s) \qquad \text{eq. 1.141}$$

Thus, the slopes of the μ(T)-plot will increase from solid to liquid to gas, and so will the y-axis intercepts (Figure 1.30).

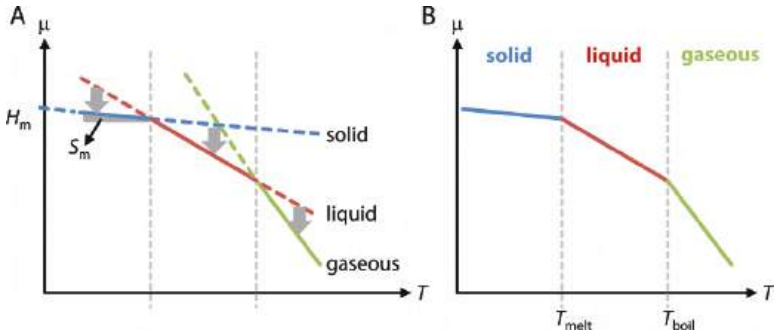


FIGURE 1.30: The chemical potential of solid, liquid, and gaseous states as a function of temperature. A: The chemical potential depends linearly on temperature. The slope of the line is the negative molar entropy $-S_m$, the intercept with the y-axis is the molar enthalpy H_m. The molar entropy S_m of a compound is low in the solid state, higher in the liquid state, and highest in the gaseous state. The molar enthalpy H_m increases in the same order. Both slopes and y-axis intercepts therefore increase from solid state (blue) to liquid state (red) to gaseous state (green). Spontaneous processes are driven by a decrease in chemical potential (gray arrows). B: Phase diagram. The most stable state is always the one with the lowest chemical potential: the solid state at low temperatures, the liquid state at intermediate temperatures, and the gas at high temperatures. At the melting temperature T_{melt}, the chemical potentials of solid and liquid are equal, and solid and liquid are in equilibrium. Similarly, liquid and gas are in equilibrium at the boiling temperature T_{boil}.

Spontaneous processes are driven by a decrease in the chemical potential. If we now look at the straight lines in Figure 1.30 that represent the chemical potential of the different states, we see that they divide the temperature range into three different regions. At low temperatures, the chemical potential of the solid is lower than the chemical potential of the liquid state. As a consequence, the transition from solid to liquid is associated with an increase in the chemical potential, and does not occur spontaneously. The reverse process, the transition from liquid to solid, however, leads to a decrease in chemical potential, and is spontaneous. In this temperature range, the solid is thus the stable phase. The lines for solid and liquid states intersect at a certain temperature. At this temperature, the chemical potentials of the solid and liquid state are equal, and the two are in equilibrium. This temperature is the melting point of the solid (or the freezing point of the liquid, depending on the starting point). The part of the system that is in the liquid state is called the liquid phase, the solid part is the solid phase. A *phase* is defined as a continuous part of a system that is homogeneous in

composition and physical state. At temperatures above the melting point, the chemical potential of the liquid state is lower than the chemical potential of the solid, and any solid present will be converted into liquid in a spontaneous reaction (with a decrease in chemical potential). At a higher temperature, the lines for the chemical potential of the liquid and gaseous states intersect, meaning the chemical potentials of these phases are equal, and they are in equilibrium. This temperature is the boiling or evaporation point of the liquid, or the condensation temperature of the gas. At temperatures above the evaporation temperature, the gas will be the phase with the lowest chemical potential, and all phase transitions toward the gas phase lead to a decrease in the chemical potential, and are therefore spontaneous. Thus, we can generalize that the state with the lowest chemical potential is the most stable state under the prevalent conditions. A diagram that shows the most stable state for a certain parameter, such as the temperature, is called a *phase diagram*. Phase diagrams are also used to describe phase transitions of biomolecules, such as protein folding and unfolding, the hybridization or dissociation of DNA strands, or phase transitions in lipid bilayers (Box 1.11).

BOX 1.11: PHASE DIAGRAMS OF BIOMOLECULES.

Instead of considering solid, liquid, and gaseous state, we can also use the chemical potential to rationalize protein folding, unfolding, and stability (Figure 1.31). Eq. 1.139 also holds for the chemical potential of the folded state of a protein, and the chemical potential of the folded state will decrease linearly with temperature (again under the assumption that H_m and S_m are temperature-independent). The folded state is more ordered than the unfolded state of a protein, and the molar entropy of the folded state, $S_{m,folded}$, will be lower than the molar entropy of the unfolded state, $S_{m,unfolded}$. To unfold a protein, interactions between side chains and main-chain atoms need to be overcome, which requires energy. The molar enthalpy of the unfolded state, $H_{m,unfolded}$, will therefore be higher than the molar enthalpy of the folded state, $H_{m,folded}$. Thus, the curve for the chemical potential of the unfolded state has a higher slope and larger intercept than the curve for the folded state. The intersection point of the two curves defines the unfolding temperature of the respective protein. The chemical potentials of folded and unfolded states are equal, and both states are in equilibrium. At temperatures below the unfolding temperature, the folded state has a higher chemical potential, and folding is the spontaneous reaction. At higher temperatures, unfolding is spontaneous. The same scheme applies to phase changes of nucleic acids (melting of double-stranded DNA to single strands) or to phase transitions in lipid membranes (transition from solid phase (gel) to liquid phase).

(*Continued*)

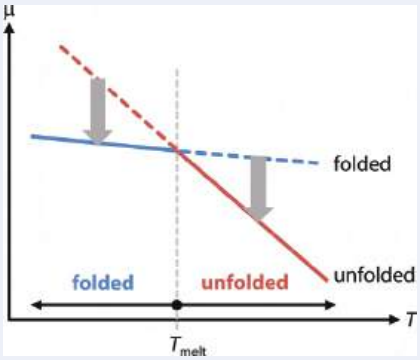


FIGURE 1.31: Phase diagram for a protein that can be heat-denatured. The chemical potential of the folded and unfolded state depends linearly on temperature, with the slope determined by the molar entropy S_m, and the *y*-axis intercept defined by the molar enthalpy H_m. The lines representing the chemical potential intersect at the melting temperature T_{melt} ($\Delta\mu = 0$, equilibrium). At temperatures below T_{melt}, the chemical potential of the unfolded state is higher, providing a driving force for folding (arrow). At temperatures above T_{melt}, the chemical potential of the folded state is higher, providing a driving force for unfolding. Hence, at lower temperatures, the folded state is more stable, at higher temperatures the unfolded state is more stable. Note that the chemical potential is only linear with temperature if H_m and S_m are temperature-independent. When H_m and S_m are temperature-dependent, plots of the chemical potential μ as a function of temperature are non-linear (see also Figure 1.26).

1.7.3 Pressure and temperature dependence of the chemical potential

In analogy to the standard free energy G^0, the standard chemical potential μ^0 of gases, liquids, and solids is defined as the chemical potential of the pure compounds under standard pressure (10^5 Pa) and at standard temperature (298.15 K). The standard chemical potential of elements is zero by definition. Pure elements hence serve as a reference state, as we have seen in the definition of enthalpies, entropies, and free energies of formation (Section 1.6; Figure 1.24).

The pressure- and temperature dependence of the chemical potential directly follows from the pressure- and temperature dependence of the free energy (eq. 1.122, eq. 1.125). The chemical potential μ at a pressure p is

$$\mu(p,T) = \mu^0(T) + RT\ln\frac{p}{p^0} \qquad \text{eq. 1.142}$$

$\mu^0(T)$ in eq. 1.142 refers to standard pressure p^0 (10^5 Pa), but does not have to be the value at standard temperature, as long as the temperature is constant. The second term corrects the chemical potential for deviations of the actual pressure p from the standard pressure p^0. The temperature dependence of the chemical potential at constant pressure is described by the Gibbs-Helmholtz equation with the molar enthalpy H_m and molar entropy S_m at the actual temperature T:

$$\mu(T) = G_m(T) = H_m(T) - T\cdot S_m(T) \qquad \text{eq. 1.139}$$

1.7.4 The chemical potential as a partial molar property

In Section 1.7 (eq. 1.129 and eq. 1.130), we introduced the chemical potential as the change in free energy caused by a change in amount of substance. In the next Section 1.7.5, we will see that the chemical potential is a very useful property to describe individual components in mixtures. The chemical potential μ_i of a component i of a mixture is defined as the change in free energy of the entire system at constant pressure and temperature when the amount of substance of the component

i is changed, while the amount of all other components remains constant. We can express this definition of the chemical potential μ_i as the partial derivative of *G* according to eq. 1.143:

$$\mu_i = \left(\frac{\partial G}{\partial n_i}\right)_{p,T,n_j \neq n_i} \qquad \text{eq. 1.143}$$

The index $p,T,n_j \neq n_i$ indicates that pressure *p* and temperature *T* as well as the amounts n_j of all other components of the mixture remain constant; the only changing parameter is n_i. The chemical potential μ_i is also called the *partial molar free energy* of the component *i*. For a pure compound, eq. 1.143 reduces to eq. 1.131.

By analogy to the definition of the chemical potential as the partial molar free energy, we can define a number of partial molar properties for individual components of mixtures, such as the partial molar enthalpy $H_{m,i}$,

$$H_{m,i} = \left(\frac{\partial H}{\partial n_i}\right)_{p,T,n_j \neq n_i} \qquad \text{eq. 1.144}$$

the partial molar entropy $S_{m,i}$,

$$S_{m,i} = \left(\frac{\partial S}{\partial n_i}\right)_{p,T,n_j \neq n_i} \qquad \text{eq. 1.145}$$

or the partial molar volume $V_{m,i}$:

$$V_{m,i} = \left(\frac{\partial V}{\partial n_i}\right)_{p,T,n_j \neq n_i} \qquad \text{eq. 1.146}$$

Partial molar properties describe how the particular state function or state variable changes with the amount of one of the components of the mixture. It is important to note that the partial molar volume for components of a mixture is not identical to the molar volume of the pure compounds. Instead, all partial molar properties depend on the composition of the mixture. As a consequence, the partial molar properties of the components of the mixture are not independent: altering the partial molar property for one component of a mixture leads to a change in the partial molar property for the other components.

We can describe the volume, entropy, enthalpy, or free energy of the entire mixture by summing the partial molar properties for each component, weighted by the amount of substance n_i of each constituent:

$$V = \sum_i n_i V_{m,i}$$

$$S = \sum_i n_i S_{m,i} \qquad \text{eq. 1.147}$$

$$H = \sum_i n_i H_{m,i}$$

$$G = \sum_i n_i G_{m,i} = \sum_i n_i \mu_i$$

The relation of molar volume, partial molar volume, and its dependence on the composition of the mixture can be illustrated considering a water-ethanol mixture as an example. The partial molar

volume of pure ethanol, $V_{m,i}$, equals its molar volume V_m. The partial molar volume of pure water is also identical to its molar volume. In mixtures of ethanol and water, their partial molar volumes deviate from the molar volume of the pure compounds. The exact value for the partial molar volumes depends on the composition of the mixture, *i.e.* on the molar fractions of ethanol and water in the mixture. Because of favorable interactions between water and ethanol, the partial molar volume of water and ethanol in the mixture is smaller than the partial molar volume of the pure substances. As a consequence, when we mix 50 mL of ethanol with 50 mL of water, we do not obtain 100 mL of an ethanol/water mixture, but a smaller volume.

1.7.5 The chemical potential of compounds in mixtures

To calculate the chemical potential for mixtures, we have to take into account the fraction of each component *i* of the mixture. We again start our consideration from ideal gases and their mixtures. According to the ideal gas equation (eq. 1.1), the pressure p of a gas is proportional to the ratio n/V of amount of substance n and the volume V, which is equal to the concentration c. Thus, the pressure of a gas is a measure of the concentration. For a mixture of gases, we can express the fraction of each component *i* by its *partial pressure* p_i. The sum of the partial pressure values p_i for all components is the pressure p of the gas mixture:

$$p = \sum_i p_i \qquad \text{eq. 1.148}$$

The additivity of partial pressures to the total pressure of a gas mixture is called *Dalton's law*. The chemical potential for the component *i* in this gas mixture is

$$\mu_i\left(p^0, T\right) = \mu_i^0\left(p^0, T\right) + RT \ln \frac{p_i}{p^0} \qquad \text{eq. 1.149}$$

μ_i^0 refers to the chemical potential of the component *i* in its standard state (pure gas) and at standard pressure. p_i is the partial pressure of component *i*, and p^0 is the standard pressure.

Instead of using partial pressures, we can also relate the chemical potential of mixtures to the *mole fraction* x_i of component *i*. The mole fraction of component *i* in a mixture of *i* different components is defined as

$$x_i = \frac{n_i}{\sum_i n_i} = \frac{n_i}{n} \qquad \text{eq. 1.150}$$

n_i is the amount of component *i*, n is the total amount of substance in the mixture. Thus, x_i is in the range of zero to one. The sum of x_i for all components is one, which is very convenient in the treatment of binary mixtures (Section 1.7.7). From the ideal gas equation (eq. 1.1), we can derive

$$x_i = \frac{n_i}{n} = \frac{p_i V}{RT} \cdot \frac{RT}{pV} = \frac{p_i}{p} \qquad \text{eq. 1.151}$$

By rearranging eq. 1.151 we obtain *Raoult's law*: in a mixture of (ideal) gases, the partial pressure p_i of each component is the total pressure p of the mixture multiplied by the mole fraction x_i of component *i*:

$$p_i = x_i \cdot p \qquad \text{eq. 1.152}$$

Combining eq. 1.149 and eq. 1.152, we can now write the concentration dependence of the chemical potential at standard pressure p^0 as

$$\mu_i = \mu_i^0\left(p^0, T\right) + RT\ln x_i \qquad \text{eq. 1.153}$$

For $x_i \to 1$, *i.e.* pure component *i*, the logarithmic term approaches zero, and the chemical potential of component *i* becomes the standard chemical potential μ_i^0, which is just the chemical potential of the pure compound *i*. The mole fraction x_i expresses the fraction of components not only in gas mixtures, but also in liquid mixtures and solutions:

$$x_i = \frac{n_i}{n} = \frac{c_i V}{cV} = \frac{c_i}{c} \qquad \text{eq. 1.154}$$

where *n* and *c* denote the total amount and concentration of the mixture. Eq. 1.153 is therefore useful for the description of gaseous and liquid mixtures as well as solutions (Section 1.7.6).

We will see in the following that we can transfer our knowledge about the chemical potential of components of gas mixtures directly to liquids and liquid mixtures (Figure 1.32). First, we consider a pure liquid A with a chemical potential μ_A^*. The star indicates that we refer to a pure compound. The liquid phase (l) is in equilibrium with the gas phase (g). As a consequence, the gas phase contains vapor of compound A with a partial pressure p_A^*. The pressure p_A^* is called the *vapor pressure* of A.

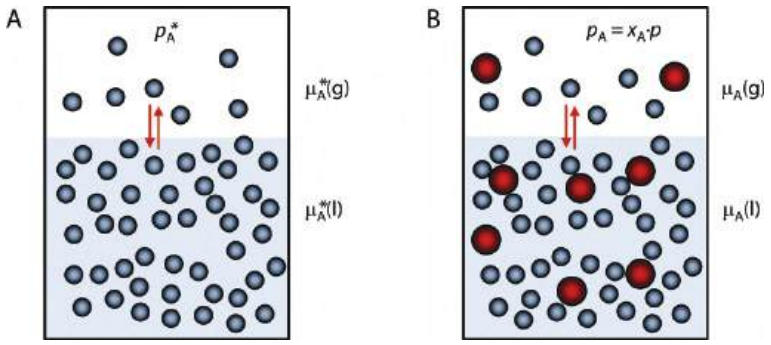


FIGURE 1.32: Vapor pressure of liquids and liquid mixtures. A: A pure liquid (blue) is in equilibrium with the gas phase (white) above. The pressure of the gas phase is defined as the vapor pressure of the liquid p_A^*. B: In a liquid mixture, all components are in equilibrium in liquid and gas phases. Their partial vapor pressure in the gas phase is proportional to the mole fraction x_i in the mixture.

Because liquid and gas phase (vapor) are in equilibrium, the chemical potential of the compound A in both phases must be equal:

$$\mu_A^*(l) = \mu_A^*(g) \qquad \text{eq. 1.155}$$

The stars in eq. 1.155 again mark properties of the pure compound. The chemical potential of A in the vapor is described by eq. 1.149. Together, we can therefore write

$$\mu_A^*(l) = \mu_A^0\left(p^0, T\right) + RT\ln\frac{p_A^*}{p^0} \qquad \text{eq. 1.156}$$

Eq. 1.156 describes the chemical potential of the pure compound A in the liquid phase. To describe the chemical potential of compound A as part of a liquid mixture, we need to express the chemical potential of the gas phase as a function of the partial pressure p_A (eq. 1.149). The chemical potential $\mu_A(l)$ of A in the liquid mixture then becomes

$$\mu_A(l) = \mu_A^0\left(p^0, T\right) + RT\ln\frac{p_A}{p^0} \qquad \text{eq. 1.157}$$

Subtraction of eq. 1.156 from eq. 1.157 gives

$$\mu_A(l) - \mu_A^*(l) = RT \ln \frac{p_A}{p^0} - RT \ln \frac{p_A^*}{p^0} \qquad \text{eq. 1.158}$$

which can be rearranged to

$$\mu_A(l) = \mu_A^*(l) + RT \ln \frac{p_A}{p_A^*} \qquad \text{eq. 1.159}$$

The chemical potential of compound A in a liquid mixture thus depends on the partial vapor pressure of A in the mixture relative to the vapor pressure of pure A. Following Raoult's law (eq. 1.152), the partial vapor pressure p_A can be expressed in terms of the mole fraction x_A and the vapor pressure of pure A, p_A^*, as

$$p_A = x_A \cdot p_A^* \qquad \text{eq. 1.160}$$

We can therefore convert eq. 1.159 into

$$\mu_A(l) = \mu_A^*(l) + RT \ln \frac{x_A \cdot p_A^*}{p_A^*} = \mu_A^*(l) + RT \ln x_A \qquad \text{eq. 1.161}$$

Eq. 1.161 corresponds to the dependence of the chemical potential on the mole fraction described in eq. 1.153. Thus, the chemical potential of components in gaseous and liquid mixtures depends on their mole fraction x_i.

Instead of using the mole fraction x_i, we can also describe the chemical potential of liquid mixtures in terms of molar concentration c by

$$\mu_i = \mu_i^0 + RT \ln \frac{c_i}{c_i^0} \qquad \text{eq. 1.162}$$

In eq. 1.162, μ_i^0 refers to the chemical potential of the compound in its standard state (pure liquid), c_i is the actual concentration of component *i*, and c_i^0 is the concentration of *i* in the standard state (pure liquid).

For solids, the chemical potential is concentration-independent. The chemical potential of each component in the solid mixture is therefore equal to the standard potential of the pure compound:

$$\mu_i = \mu_i^0 \qquad \text{eq. 1.163}$$

We will briefly revisit the chemical potential of solids in the context of electrochemistry (Chapter 5). The majority of biochemical reactions occur in solutions, however, and in the following Sections, we therefore focus on the chemical potential and properties of solutions.

1.7.6 The chemical potential of solutions

Solutions are a special case of liquid mixtures in which one component, the *solvent*, is in large excess over the other components, the *solutes*. We can express the chemical potential of solvent and solutes by eq. 1.162:

$$\mu_i = \mu_i^0 + RT \ln \frac{c_i}{c_i^0} \qquad \text{eq. 1.162}$$

Although the overall form of the chemical potential is identical for solvent and solute(s), the relevant concentrations c_i^0 in their standard states are different. By definition, the solvent is present in large

excess, and the solute is the minor component. Therefore, the standard state for the solvent is the pure liquid, whereas for solutes, the concentration in the standard state is defined as 1 M (Section 1.6.3). For aqueous solutions, we have to use a standard-state concentration for the solvent water (55.5 M), and 1 M for the solutes. We have seen before that the biochemical standard state defines a different reference state for protons in solutions (pH 7, $c^0 = 10^{-7}$ M). It is important to use a consistent system of standard states, otherwise the calculated values will not match (Box 1.12)!

BOX 1.12: THE CHEMICAL POTENTIAL OF PROTONS IN AQUEOUS SOLUTION.

The chemical potential for H^+ in a 0.1 M solution of HCl can be calculated by using the chemical standard state (1 M H^+, corresponding to a pH of 0) and the corresponding standard state chemical potential:

$$\mu_{H^+} = \mu^0_{H^+} + RT \ln \frac{0.1M}{1M}$$

Alternatively, we can use the biochemical standard state (pH 7 = 10^{-7} M for protons), and calculate the chemical potential as

$$\mu_{H^+} = \mu^{0'}_{H^+} + RT \ln \frac{0.1M}{10^{-7}M}$$

using the standard chemical potential at pH 7 (indicated by the prime) as a reference. Because the chemical potential is a state function, the result of both calculations must be identical, but the reference values μ^0 and $\mu^{0'}$ differ.

We can also describe the chemical potential of solvent and solutes in terms of the mole fraction x_i using eq. 1.153.

$$\mu_i = \mu^0_i + RT \ln x_i \qquad \text{eq. 1.153}$$

For a dilute solution, the mole fraction is near unity for the solvent, and very small for the solutes. The reference chemical potential μ_i^0 (p,T) now refers to the chemical potential of the pure compound ($x_i = 1$). We will see in the following Section 1.7.7 that the description of the chemical potential of solutions in terms of the mole fraction is very practical, and enables us to quantitatively describe important properties of solutions.

The general equation for the mole fraction dependence of the chemical potential of a solution only holds for *ideal solutions*. In an ideal solution, the molecules of the solute and the solvent do not interact among themselves or with each other. This is the exception, however, and in most solutions, interactions between solvent and solute do occur. Nevertheless, many dilute solutions ($x_{solute} \ll 1$) behave as ideal solutions, and their behavior can be described by eq. 1.153. Solutions in biological systems, however, frequently deviate from ideality because the solutes interact with the solvent, or the solutes interact with each other, and they are highly concentrated. For such *non-ideal solutions*, the chemical potential does not follow eq. 1.153:

$$\mu_i \neq \mu^0_i + RT \ln x_i \qquad \text{eq. 1.164}$$

We can convert the inequality into an equation, however, by correcting the mole fraction to a value that corresponds to the effective mole fraction a_i, *i.e.* a mole fraction that under ideal conditions would lead to the real chemical potential of the solution (Figure 1.33):

$$\mu_i = \mu_i^0 + RT \ln a_i \qquad \text{eq. 1.165}$$

The effective mole fraction a_i is called the *activity* of the component *i*. The activity can be expressed as the actual mole fraction x_i, multiplied by an *activity coefficient* γ_i:

$$a_i = x_i \cdot \gamma_i \qquad \text{eq. 1.166}$$

By combining eq. 1.165 and eq. 1.166, we obtain

$$\mu_i = \mu_i^0 + RT \ln x_i \gamma_i \qquad \text{eq. 1.167}$$

for the chemical potential of a non-ideal solution. Eq. 1.167 can also be written as

$$\mu_i = \mu_i^0 + RT \ln x_i + RT \ln \gamma_i \qquad \text{eq. 1.168}$$

The first two terms in eq. 1.168 correspond to the chemical potential of an ideal solution with a mole fraction x_i, while the last term corrects the chemical potential for non-ideal effects. The chemical potential of the non-ideal solution is thus just the chemical potential of an ideal solution, corrected by $RT \ln\gamma_i$ to account for non-ideality:

$$\mu_i = \mu_{i,\text{ideal}} + RT \ln \gamma_i \qquad \text{eq. 1.169}$$

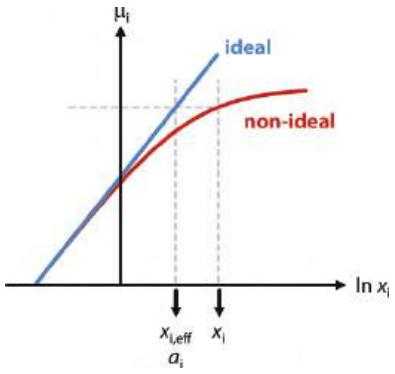


FIGURE 1.33: Dependence of the chemical potential on the mole fraction for ideal and non-ideal solutions. In an ideal solution, the chemical potential of the solvent depends linearly on the logarithm of the mole fraction of the solute, ln x_i (blue). The chemical potential of a non-ideal solution deviates from this linear dependence (red). The effective mole fraction $x_{i,\text{eff}}$ (activity a_i) of a non-ideal solution corresponds to the mole fraction of an ideal solution with the same chemical potential.

1.7.7 Colligative properties

When we dissolve a solute in a pure solvent A, the mole fraction x_A of the solvent decreases, and its chemical potential μ_A changes (eq. 1.153). The chemical potential μ_A of the solvent A is

$$\mu_A = \mu_A^0 + RT \ln x_A \qquad \text{eq. 1.170}$$

For a pure solvent ($x_A \to 1$), the chemical potential is μ_A^0, the chemical potential at standard conditions, *i.e.* the pure liquid. If we now consider a solution of the solute B in solvent A, we can express the mole fraction x_A of the solvent by the mole fraction x_B of the solute as $(1-x_B)$. Eq. 1.170 then becomes

$$\mu_A = \mu_A^0 + RT \ln(1 - x_B) \qquad \text{eq. 1.171}$$

For a very dilute solution ($x_B \ll 1$), we can approximate the logarithm as

$$\ln(1 - x_B) \approx -x_B \qquad \text{eq. 1.172}$$

and obtain

$$\mu_A = \mu_A^0 - x_B RT \qquad \text{eq. 1.173}$$

The solute thus reduces the chemical potential of the solvent. Strikingly, this effect only depends on the mole fraction of the solute, but is independent of its chemical identity. Properties that only depend on the numbers of particles involved, but not on their identity, are called *colligative properties*.

If we now consider the effect of the solute on the chemical potential of the solvent as a function of temperature, we see that colligative properties have an important consequence: at any given temperature, the chemical potential of the solvent in a solution will be lower than that of the pure solvent (Figure 1.34). The temperature at which the chemical potentials of solid and liquid states are equal, the melting point of the solid or freezing point of the solution, therefore shifts to lower temperatures (*freezing point depression*). Conversely, the temperature at which the chemical potentials of liquid and gaseous states are equal, the evaporation temperature, shifts to higher temperatures (*boiling point elevation*).

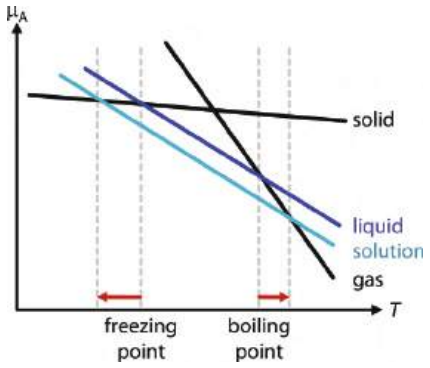


FIGURE 1.34: Effect of solute on the freezing and boiling points of the solvent. The solute decreases the chemical potential of the solvent relative to the pure solvent at all temperatures (blue: liquid, cyan: solution). The chemical potential of solid or gaseous phases (black) is not affected. As a consequence, the intersection point for the chemical potential of the solid phase and the solution is shifted to lower temperatures: the freezing point (melting point of the solid phase) is decreased compared to the pure liquid. The intersection point for the chemical potential of the solution and of the gas phase is higher, the boiling point is increased compared to the pure solvent.

The change in boiling point caused by the solute can be calculated using the chemical potential of the solvent in liquid and gaseous phases. At the evaporation temperature, the liquid and gaseous phases of the solvent are in equilibrium, and we can write:

$$\mu_A(g) = \mu_A^0(l) + RT \ln x_A \qquad \text{eq. 1.174}$$

By rearranging to

$$\frac{\mu_A(g)-\mu_A^0(l)}{RT}=\ln x_A \qquad \text{eq. 1.175}$$

and by substitution of the chemical potential by the change in (molar) free energy associated with the phase transition, ΔG_{tr}, and for ΔH_{tr} and ΔS_{tr} using the Gibbs-Helmholtz equation (eq. 1.111), we obtain

$$\frac{\Delta G_{tr}}{RT}=\frac{\Delta H_{tr}-T\Delta S_{tr}}{RT}=\ln x_A \qquad \text{eq. 1.176}$$

In equilibrium, ΔG_{tr} becomes zero (Section 1.6). From the Gibbs-Helmholtz equation it follows that ΔS_{tr} can be substituted by $\Delta H_{tr}/T_{tr}$, which gives

$$\frac{\Delta H_{tr}}{RT}-\frac{T\Delta H_{tr}}{T_{tr}\cdot RT}=\ln x_A \qquad \text{eq. 1.177}$$

Eq. 1.177 can be rearranged to

$$\frac{\Delta H_{tr}}{R}\left(\frac{1}{T}-\frac{1}{T_{tr}}\right)=\ln x_A \qquad \text{eq. 1.178}$$

The term in parentheses can be expanded to

$$\frac{T_{tr}-T}{T\cdot T_{tr}}$$

For small changes in the transition temperature ($T \approx T_{tr}$), this term can be approximated as

$$\frac{T_{tr}-T}{T\cdot T_{tr}}\approx-\frac{\Delta T}{T_{tr}^2} \qquad \text{eq. 1.179}$$

ΔT refers to the shift in the transition temperature (T–T_{tr}) caused by the solute. On the right-hand side of eq. 1.178, the mole fraction of the solvent A can be expressed as (1–x_B):

$$-\frac{\Delta H_{tr}}{R}\cdot\frac{\Delta T}{T_{tr}^2}=\ln x_A=\ln(1-x_B) \qquad \text{eq. 1.180}$$

For small x_B, *i.e.* dilute solutions of B in A, we can again approximate ln (1–x_B) by –x_B and obtain

$$-\frac{\Delta H_{tr}}{R}\cdot\frac{\Delta T}{T_{tr}^2}\approx-x_B \qquad \text{eq. 1.181}$$

Rearranging eq. 1.181 yields the shift ΔT in the transition temperature:

$$\Delta T\approx\frac{x_B\cdot T_{tr}^2\cdot R}{\Delta H_{tr}} \qquad \text{eq. 1.182}$$

The shift in evaporation temperature by the solute thus depends on the mole fraction of the solute, x_B, and not on its identity: the boiling point increase by solutes is a colligative property. We can perform the same calculation to determine the shift in freezing point/melting point, which has an opposite sign (melting point reduction, see Figure 1.34). Due to the stronger dependence of the gas chemical potential on the temperature compared to the solid, and the higher slope of the corresponding μ(T) curves, the melting point reduction is always more pronounced than the boiling point increase (Figure 1.34).

As a consequence of the relationship between ΔT and x_B (eq. 1.182), we can measure the shift in the transition temperature, ΔT, and determine the mole fraction x_B, provided that the associated transition enthalpy ΔH_{tr} is known. An instrument that measures the boiling point elevation is called an *ebullioscope*. A *cryoscope* measures the depression of the freezing point. Ebullioscopy can be used to determine the alcohol content of liquids from the boiling point of the mixture. Measuring the melting temperature T_m of a compound is a criterion for purity. If the melting point of a supposedly pure compound is lower than the reference value for the pure compound, impurities are present.

Osmosis, a phenomenon caused by diffusion through semipermeable membranes, is a biologically important example of a colligative property. Biological membranes act as selective barriers because they are not permeable for most solutes but can be readily passed by water. Membranes that are permeable for the solvent, in biological systems water, but not for the solute, are called *semipermeable* (Figure 1.35). If the solute concentration is high inside a membrane-limited compartment, but low outside, the chemical potential of water (the solvent) is higher on the outside than on the inside of the membrane. The higher chemical potential is a driving force for flow, and water will flow into the compartment to equilibrate the chemical potential. However, as the solute cannot pass the membrane, its concentration will always remain higher inside. No matter how much water flows into the cell, the equilibrium situation cannot be reached. Unlimited influx will eventually lead to bursting of the membrane. This is the reason why heavy rains in fall are dreaded by winemakers: the grapes have reached a high concentration of sugar, and the rain is almost pure water. Therefore, water will penetrate the skin of the grapes, causing them to burst and destroying the harvest.

We can also imagine the reverse situation, with pure water (or a dilute solution) inside the membrane-limited compartment, and a solution (or a solution of higher concentration) on the outside. In this case, the higher chemical potential of water in the inside will lead to uncontrolled efflux of water, called *plasmolysis*. We have all experienced plasmolysis when dressing lettuce: the higher salt concentration of the dressing compared to the interior of the leaves' cells leads to an efflux of water, and the previously crisp leaves turn limp if we add the dressing too early.

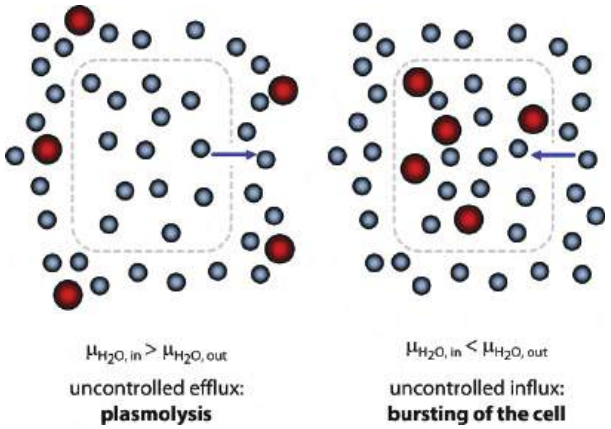


FIGURE 1.35: Diffusion through a semipermeable membrane. Left: If the salt concentration (red) is higher outside the membrane-limited compartment, the chemical potential μ of water (blue) is higher on the inside than on the outside. Water spontaneously flows from the inside to the outside (blue arrow) to reduce the difference in chemical potential (plasmolysis). Right: If the salt concentration is higher inside the compartment, the higher chemical potential of water outside will drive influx of water (blue arrow), eventually leading to a bursting of the membrane.

Due to the limited elasticity of cellular membranes, the influx of water into the cell will cause a pressure π to build up (Figure 1.36). This *osmotic pressure* will prevent further influx of water, and equilibrium is reached.

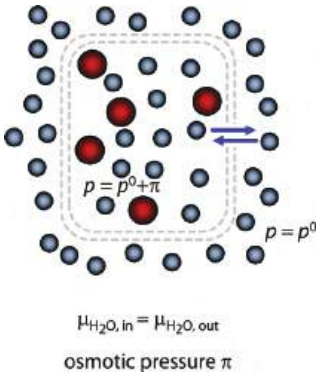


FIGURE 1.36: Osmosis. A cell is separated from the exterior by a biological membrane, a lipid bilayer (gray broken lines). Inside, the solute (red) concentration is higher than outside. The higher chemical potential μ of water (blue) outside acts as a driving force for influx of water into the cell (blue arrow). Because of the limited elasticity of the membrane, the pressure inside will increase with the influx of water, and this counteracts further influx of water. The pressure difference between inside and outside in equilibrium is called the osmotic pressure π.

We can express the pressure- and mole fraction-dependent chemical potential of the solvent A in the solution inside the membrane-limited compartment as

$$\mu_A = \mu_A^0 + \int_{p^0}^{p^0+\pi} V_{m,A}\,dp + RT\ln x_A \qquad \text{eq. 1.183}$$

The chemical potential of the pure solvent on the outside is simply the standard potential μ_A^0. In equilibrium, the chemical potentials outside and inside are equal:

$$\mu_A^0 + \int_{p^0}^{p^0+\pi} V_{m,A}\,dp + RT\ln x_A = \mu_A^0 \qquad \text{eq. 1.184}$$

We can now evaluate the integral

$$\int_{p^0}^{p^0+\pi} V_{m,A}\,dp = \pi \cdot V_{m,A} \qquad \text{eq. 1.185}$$

and obtain

$$\pi \cdot V_{m,A} = -RT\ln x_A \qquad \text{eq. 1.186}$$

Substituting $1-x_B$ for x_A yields

$$\pi \cdot V_{m,A} = -RT\ln(1-x_B) \qquad \text{eq. 1.187}$$

Again, for dilute solutions and $x_B \ll 1$ we can approximate the logarithm as $-x_B$ and obtain

$$\pi \cdot V_{m,A} \approx RTx_B \qquad \text{eq. 1.188}$$

Rearrangement and substitution of x_B by n_B/n and $V_{m,A}$ by V/n yields

$$\pi = \frac{x_B}{V_{m,A}} \cdot RT = \frac{n_B}{n} \cdot \frac{n}{V} \cdot RT \qquad \text{eq. 1.189}$$

and

$$\pi = \frac{n_B}{V} RT = c_B RT \qquad \text{eq. 1.190}$$

According to eq. 1.190, the osmotic pressure π that builds up depends on the concentration or mole fraction of the solute B, but not on its identity. The effect of osmosis thus also is a colligative property. The osmotic effect can be used to determine concentrations. It is also exploited to exchange buffers by dialysis (Box 1.13), and is the principle behind dialysis for patients with kidney problems. Equilibrium dialysis can be used to determine equilibrium constants (Section 2.1; Box 1.13).

BOX 1.13: DIALYSIS.

Dialysis is used for buffer exchange of protein solutions (Figure 1.37). The protein solution is filled into a dialysis tube that is permeable for water and small ions, but not for the protein. The dialysis tube is placed into a large reservoir of the target buffer. Osmosis leads to equilibration of the ion and water chemical potentials across the membrane. Because of the large difference in volume of the solutions inside the dialysis tube and outside, this will lead to a near-equilibration of ion concentrations inside and outside, and effectively to an "exchange" of the buffer in which the protein is dissolved. Note that the osmotic pressure should be considered when setting up the dialysis: if the salt concentration is much higher inside than outside, the water flow into the dialysis bag can cause it to burst! Bursting can be avoided by leaving some void volume within the bag.

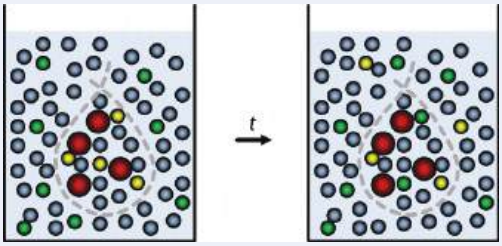


FIGURE 1.37: Dialysis for buffer exchange. A macromolecule in starting buffer (water, blue spheres, and salt 1, yellow spheres) is filled into a dialysis tube, and placed into a large reservoir of the desired buffer (not to scale, water and salt 2, green spheres). Water and ions can freely diffuse between the inside and outside of the tube, and concentrations will equilibrate. The reservoir is much larger than depicted, leading to (almost) equal salt concentrations in equilibrium and thus "exchange" of buffers. At the end of the dialysis, the macromolecule will effectively be dissolved in the desired buffer.

Equilibrium dialysis can be used to determine dissociation constants of macromolecule-ligand complexes (Section 2.2). A macromolecule solution of known concentration is filled into one chamber, and the ligand into the second chamber, separated by a semipermeable membrane (Figure 1.38). The ligand can pass the membrane, but the macromolecule cannot and is confined to chamber one. The equilibrium is reached when the chemical potential of the ligand on both sides of the membrane is equal or, more specifically, when the chemical potential of the free ligand in both chambers is equal. The bound molecules do not contribute to the chemical potential of the ligand, and the equilibrium condition is

(*Continued*)

$$\mu^{\text{left}}_{\text{free ligand}} = \mu^{\text{right}}_{\text{ligand}} \qquad \text{eq. 1.191}$$

Eq. 1.191 implies that the concentration of the free ligand in the left chamber is equal to the ligand concentration in the right chamber. Thus, if we determine the total concentration of ligand in both chambers, $c_{\text{L,right}}$ (= $c_{\text{L,left,free}}$), and $c_{\text{L,left,tot}}$, we can calculate the concentration of the macromolecule-ligand complex, $c_{\text{L,left,bound}}$, from the difference $c_{\text{L,left,tot}} - c_{\text{L,right}}$ and obtain K_d (Section 2.2) as

$$K_d = \frac{[M][L]}{[ML]} = \frac{\left([M]_{\text{left,tot}} - \left(c_{\text{L,left,tot}} - c_{\text{L,right}}\right)\right) \cdot c_{\text{L,right}}}{c_{\text{L,left,tot}} - c_{\text{L,right}}} \qquad \text{eq. 1.192}$$

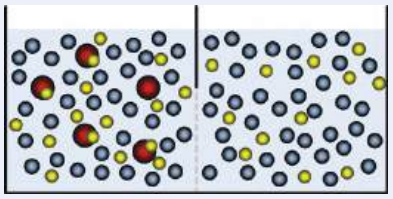


FIGURE 1.38: Equilibrium dialysis. The solution of a ligand (yellow, blue: water) for a macromolecule (red) is filled into two chambers separated by a semipermeable membrane (gray). Equilibrium is reached when the concentration of free ligand is equal in both compartments. By determining the total concentrations of the ligand in the left chamber and the ligand concentration in the right chamber, we can calculate the concentration of macromolecule-bound ligand, and the equilibrium constant K_d.

In *reverse osmosis* (Figure 1.39), external pressure is applied to concentrate a solute, or to desalinate water. The applied pressure provides the energy for the transport of water against a concentration gradient toward higher chemical potential (Section 4.3). The same principle is applied, *e.g.* to remove ethanol from alcoholic beverages. Here, ethanol and water permeate the membrane when pressure is applied, whereas larger molecule responsible for the taste cannot pass and are retained. Ethanol is removed from the permeate by distillation. Mixing of the permeate after distillation with the retentate of the reverse osmosis then generates the (almost) alcohol-free beverage.

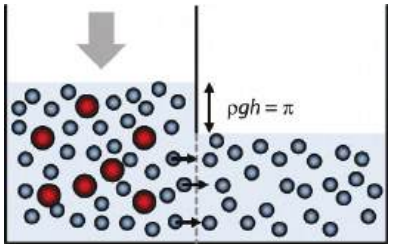


FIGURE 1.39: Reverse osmosis. A solution (left) is separated from pure water (right) by a semipermeable membrane (broken gray line) that is permeable for water (blue), but impermeable for the solute (red). The osmotic effect causes a water flow from right to left until the hydrostatic pressure of the water column ρgh equals the osmotic pressure π. Applying external pressure to the side of the solution (gray arrow) provides energy for water to pass the membrane (black arrows) against the concentration gradient, leading to a concentration of the solution, and an increase in pure solvent on the right.

All living organisms consist of cells that are delimited by membranes. It is important for organisms to control salt concentrations in the cytoplasm and the water flow across membranes to be able to counteract osmotic effects (osmoregulation, Box 1.14).

BOX 1.14: OSMOREGULATION.

Some organisms live in high-salt environments. Bacteria that thrive under high ionic strength conditions, termed halophiles, take up salts from their environment. Their complete metabolism is adapted to high ionic strength, and these organisms cannot survive when the ionic strength is reduced. Halotolerant bacteria produce osmotically active organic compounds, such as amino acids, polyalcohols, or betaines, when the external salt concentration increases. These low molecular mass compounds, also called osmoprotectants, help maintain a high osmotic value of the cytoplasm, and prevent constant efflux of water and plasmolysis.

Osmoregulation is also important for fresh-water and salt-water fish. The bodies of fresh-water fish have a higher salt concentration than the surroundings, they are *hypertonic*. Therefore, they continually take up water through their gills, and ions diffuse to the outside along the concentration gradient. Fresh-water fish have to take up ions actively, and rid themselves of water by producing highly diluted urine. Salt-water fish, in contrast, have a lower salt concentration in their body fluids than the surrounding ocean, they are *hypotonic*. To avoid constant loss of water and influx of ions, they take up salt water, and actively excrete ions.

Higher organisms excrete salts by specialized glands. The Galapagos lizards live in a high-salt environment: on the islands, there is virtually no freshwater available, and the lizards quench their thirst with amply available salt water. To eliminate the excess of salt from their bodies, they use nasal glands from which they expel highly concentrated salt drops at short intervals in a movement that very much looks like sneezing. Mangrove plants also secrete salt from glands. In vertebrates, osmoregulation is achieved by control of water excretion through the kidneys.

QUESTIONS

1.1 What are open, closed, or isolated systems? (1) a protein in solution in a sealed test tube, (2) the lungs, (3) a bird, (4) a lake, (5) a biochemical reaction in a reaction tube, (6) coffee in a thermos flask, (7) the sample chamber of a differential scanning calorimeter, (8) the sample chamber of an isothermal titration calorimeter, (9) a cuvette in a photometer (no lid).

1.2 What is the sign of the transferred energy from the perspective of the system and the surroundings for
(1) a swimmer in a pool with cold water?
(2) a heated house?
(3) a coffee cup?
(4) an ice cold drink in a warm bar?
(5) a cold hand shaking a warm hand?

1.3 Is the earth an isolated, closed, or open system?

1.4 A triathlete performs 1000 kJ of work and loses 400 kJ of heat while swimming in cold water. How does the internal energy of the triathlete change? How much does the pool (50 m length, 20 m width, 2 m depth) warm up? C_p (H_2O) = 4.18 kJ kg^{-1} K^{-1}, ρ (H_2O) = 1 g cm^{-3}.

1.5 Isothermal expansion of an ideal gas does not change its internal energy. What is the associated change in enthalpy?

1.6 During normal breathing, humans exchange about 0.5 L of air in their lungs. Calculate the amount of work that is performed during one exhalation against atmospheric pressure. How

many grams of ATP need to be hydrolyzed for breathing during a day (30 breathing cycles per minute, $\Delta G^{0'}$ (ATP hydrolysis) = –31 kJ mol^{-1}, M = 507.18 g mol^{-1}). A 100 g bar of chocolate provides about 500 kJ of energy. How long does the energy last just for breathing?

1.7 Air has a CO_2 content of about 0.05%. What is the partial pressure of CO_2 in air at $p = 1{\cdot}10^5$ Pa (1 bar)?

1.8 A diver grabs a bottle of compressed air. The pressure gauge states $p = 200{\cdot}10^5$ Pa (200 bar). Just after the descent into the dive, the pressure is reduced to $190{\cdot}10^5$ Pa. Why? The diver has directly descended to 30 m depth. What are the pressure and the temperature of the surrounding water? $T_{air,surface}$ = 30°C.

1.9 O_2 becomes toxic to humans at pressures > $1.2{\cdot}10^5$ Pa. What is the maximum depth a diver can safely dive to using air (21% O_2), assuming constant temperature? A deep-sea diver wants to dive to 200 m on a special gas mixture (Trimix). What is the maximum percentage of O_2 that this mixture may have to avoid toxic effects at depth?

1.10 The diver breathes enriched air that contains 32% (m/m_{tot}) oxygen and 68% (m/m_{tot}) nitrogen (at $p = p^0 = 10^5$ Pa and $T = T_0$ = 25°C). A partial oxygen pressure of > 1.6 10^5 Pa is lethal for humans. Is it safe for the diver to dive down to the seabed at 40 m? The diver starts with a 15 L tank of enriched air at $p = p^0$. The air needed during descent and ascent can be neglected. How long can the diver stay at 45 m with his air supply for 0.5 L breathing volume and 30 breathing cycles min^{-1}?

1.11 On a winter day (T = –10°C) you adjust the pressure of your car tires to $1.8{\cdot}10^5$ Pa. What is the pressure in summer (T = 30°C)?

1.12 The enthalpy change for the reaction of glucose ($C_6H_{12}O_6$) to CO_2 and H_2O is –2800 kJ mol^{-1}, for the reaction of ethanol to water it is –1370 kJ mol^{-1}. What is the enthalpy change during fermentation of glucose to ethanol? Is fermentation a useful metabolic pathway?

1.13 The heat capacity of water (C_p = 75 J mol^{-1} K^{-1}) is much higher than the heat capacity of air (C_p = 20 J mol^{-1} K^{-1}). Calculate the temperature change when 10 kJ of heat is transferred to 1 m^3 of water or air. $\rho(H_2O)$ = 1 g cm^{-3}; ρ(air) = 1.2 mg cm^{-3}.

1.14 Calculate the change in entropy for the conversion of 1 mol ice (H_2O(s), 0°C) into vapor (H_2O(g), 100°C) at constant pressure. Sublimation enthalpy ΔH_{subl} = 47 kJ mol^{-1}, vaporization enthalpy ΔH_{vap} = 41 kJ mol^{-1}, C_p = 75 J mol^{-1} K^{-1}.

1.15 Calculate the entropy change when 200 kJ of heat are transferred to water at 0°C and 100°C (isothermal conditions). Explain the difference from the statistical interpretation of entropy.

1.16 Our metabolism generates heat of about 100 J s^{-1}. Calculate the change in entropy of the surrounding per hour at 25°C.

1.17 0.1 mol of a reactant react to products in an isobaric reaction at 25°C and 10^5 Pa. During the reaction, 20 kJ of heat are transferred to the surroundings. Calculate ΔG^0, ΔH^0, and ΔS^0 for this process.

1.18 Calculate the reaction enthalpy, entropy, and free energy, ΔH^0_r, ΔS^0_r, and ΔG^0_r, for the amidation of glutamate to glutamine (1) at standard temperature, (2) at 37°C, and (3) at 75°C (a) taking into account the temperature dependence of ΔH and ΔS, and (b) assuming that ΔH and ΔS are temperature-independent. What can you conclude about the importance of C_p?

What does the result mean for organisms that live at moderate temperature (37°C; mesophilic organisms) and thermophilic organisms that live at 75°C?
Glutamine: $\Delta H^0_f = -826$ kJ mol^{-1}, $C_{p,m} = 184$ J mol^{-1} K^{-1}, $S_m = 195$ J mol^{-1} K^{-1}
Glutamate: $\Delta H^0_f = -1010$ kJ mol^{-1}, $C_{p,m} = 175$ J mol^{-1} K^{-1}, $S_m = 188$ J mol^{-1} K^{-1}
H_2O: $\Delta H^0_f = -290$ kJ mol^{-1}, $C_{p,m} = 75$ J mol^{-1} K^{-1}, $S_m = 70$ J mol^{-1} K^{-1}
NH_4^+: $\Delta H^0_f = -133$ kJ mol^{-1}, $C_{p,m} = 80$ J mol^{-1} K^{-1}, $S_m = 113$ J mol^{-1} K^{-1}

1.19 The energy of one photon is $E = h \cdot \nu = hc/\lambda$. How many photons of light with $\lambda = 680$ nm have to be absorbed to synthesize an ATP molecule (degree of efficiency 100%)? In what wavelength range is one photon sufficient for the synthesis of two ATP molecules? $\Delta G^{0'}$ of ATP synthesis: –31 kJ mol^{-1}.

1.20 Protein-protein interactions reduce the number of particles from two to one, which corresponds to a decrease in entropy. Why can the interaction nevertheless be entropically favorable and under what conditions?

1.21 How much does an ionic interaction between two oppositely charged side chains contribute to protein stability at the surface and in the interior of the protein?

1.22 You mix 10 mL glycerol and 90 mL water to obtain a 10% glycerol solution. The density of the mixture is $\rho_{mix} = 1.02567$ g cm^{-3}. What are the mole fraction of glycerol and the volume of the mixture? What is the reason for the volume change? What can you conclude for the necessity to take volume changes into account when stabilizing proteins by using 10% glycerol buffers? M(glycerol) = 92.09 g mol^{-1}, $M(H_2O)$ = 18 g mol^{-1}, ρ(glycerol) = 1.25802 g cm^{-3}, $\rho(H_2O)$ = 0.99708 g cm^{-3}.

1.23 Calculate the difference between μ^0 and $\mu^{0'}$ for protons in aqueous solutions.

1.24 A membrane transporter exports H^+ ions from the cytoplasm of a soil bacterium into the environment to maintain the pH of the cytoplasm at pH 7.5. The pH of the soil is 7.8, but drops to 7.2 (6.4) after acidic rain. When the gardener brings out fertilizer, it increases to 8.2. What is the effect on the energy demand (in the form of ATP hydrolysis, $\Delta G = -31$ kJ mol^{-1}) for H^+ export? $T = 298$ K.

1.25 Is mixing of two liquids to an ideal solution a spontaneous process?

1.26 A polysaccharide solution ($c(PS)$ = 10 g L^{-1} in H_2O) has an osmotic pressure of $5 \cdot 10^3$ N m^{-2}. What is the molar mass of the polysaccharide? What is the vapor pressure of the solution compared to pure water?

1.27 Giant sequoia trees reach a height of more than 100 m. They have to transport water into the top of their crown. Can this be explained by the osmotic pressure due to the solutes in the cytoplasm? $\rho(H_2O)$ = 1 g cm^{-3}.

1.28 What concentration of NaCl is required to prevent ice formation at $T = -1$°C and at $T = -5$°C? ΔH_{melt} (H_2O) = 6 kJ mol^{-1}. The solubility of NaCl in H_2O is 359 g L^{-1}, the molar mass M is 58.44 g mol^{-1}. What is the maximum reduction in freezing point that can be achieved by a saturated NaCl solution?

REFERENCES

Ellis, R. J. (2001) Macromolecular crowding: Obvious but underappreciated. *Trends Biochem Sci* **26**(10): 597–604.
· review about thermodynamic consequences of molecular crowding for cellular processes

Luby-Phelps, K. (2013) The physical chemistry of cytoplasm and its influence on cell function: An update. *Mol Biol Cell* **24**(17): 2593–2596.
· review about molecular crowding and its impact on cellular processes

Makhatadze, G. I. and Privalov, P. L. (1993) Contribution of hydration to protein folding thermodynamics. I. The enthalpy of hydration. *J Mol Biol* **232**(2): 639–659.
· effect of the hydration of polar and non-polar groups during unfolding

Makhatadze, G. I. and Privalov, P. L. (1993) Contribution of hydration to protein folding thermodynamics. II. The entropy and Gibbs energy of hydration. *J Mol Biol* **232**(2): 660–679.
· effect of the hydration of polar and non-polar groups during unfolding

Minton, A. P. (2006) Macromolecular crowding. *Curr Biol* **16**(8): R269–271.
· introductory note on molecular crowding in the cell

Privalov, P. L. (1990) Cold denaturation of proteins. *Crit Rev Biochem Mol Biol* **25**(4): 281–305.
· review about the experimental evidence for cold denaturation of proteins

2 Energetics and chemical equilibria

2.1 THE FREE ENERGY CHANGE AND THE EQUILIBRIUM CONSTANT

We have seen in Section 1.7.1 that the change in free enthalpy, ΔG, during a reaction according to Scheme 2.1

$$aA + bB \rightleftharpoons cC + dD \qquad \text{scheme 2.1}$$

can be written as

$$\Delta G = \sum_i \mu_i \Delta n_i = \mu_C \Delta n_C + \mu_D \Delta n_D - \mu_A \Delta n_A - \mu_B \Delta n_B \qquad \text{eq. 2.1}$$

The relationship in eq. 2.1 does not only hold for irreversible reactions that proceed completely from A and B to C and D, but also for reversible reactions in which C and D can react to A and B in a reverse reaction, and the final state is characterized by equilibrium concentrations of A, B, C, and D. We can substitute the chemical potential for each compound in eq. 2.1 by the standard chemical potential μ^0 at standard concentrations of all compounds at pressure p and temperature T, plus the concentration term. For Δn_i, we use the stoichiometric coefficients in Scheme 2.1:

$$\Delta G = c\mu_C^0 + d\mu_D^0 - a\mu_A^0 - b\mu_B^0 + RT \ln \frac{\left(\frac{c_C}{c_C^0}\right)^c \left(\frac{c_D}{c_D^0}\right)^d}{\left(\frac{c_A}{c_A^0}\right)^a \left(\frac{c_B}{c_B^0}\right)^b} \qquad \text{eq. 2.2}$$

In eq. 2.2, the standard terms are summarized at the beginning, and the concentration-dependent terms are summarized in the logarithmic term at the end. Each term in the argument of the logarithm contains the actual concentration (c_A, c_B, c_C, c_D), relative to the standard concentration c_A^0, c_B^0, c_C^0, and c_D^0 of the respective compound. The ratio of the actual concentration and the standard concentration

is dimensionless, and the argument of the logarithm is also a dimensionless number. Biochemical reactions typically take place in solution, and the standard concentration of the dissolved reactants is 1 M. Division of the actual concentration by 1 M does not change the numerical value, and eq. 2.2 is therefore often written in a simplified form as

$$\Delta G = c\mu_C^0 + d\mu_D^0 - a\mu_A^0 - b\mu_B^0 + RT\ln\frac{[C]^c[D]^d}{[A]^a[B]^b} \qquad \text{eq. 2.3}$$

It is important to note that the values $[A]$, $[B]$, $[C]$, and $[D]$ in eq. 2.3 now denote the concentration relative to the standard concentration, and are dimensionless.

We can summarize all terms containing standard chemical potentials as ΔG^0, and obtain

$$\Delta G = \Delta G^0 + RT\ln\frac{[C]^c[D]^d}{[A]^a[B]^b} \qquad \text{eq. 2.4}$$

Eq. 2.4 allows calculating the free energy change ΔG involved in a reaction for any given concentration of reaction partners. When using this relationship, we always have to keep in mind that the concentrations are relative to the standard concentration.

In eq. 2.4, the activity coefficients γ of the compounds A, B, C, and D have been neglected, and the equation therefore only holds for ideal solutions. For a non-ideal solution with $\gamma \neq 1$, we have to multiply the actual concentrations of A, B, C, and D by their activity coefficients γ_A, γ_B, γ_C, and γ_D, which gives

$$\Delta G = \Delta G^0 + RT\ln\frac{([C]\gamma_C)^c([D]\gamma_D)^d}{([A]\gamma_A)^a([B]\gamma_B)^b} \qquad \text{eq. 2.5}$$

or

$$\Delta G = \Delta G^0 + RT\ln\frac{[C]^c[D]^d}{[A]^a[B]^b} + RT\ln\frac{(\gamma_C)^c(\gamma_D)^d}{(\gamma_A)^a(\gamma_B)^b} \qquad \text{eq. 2.6}$$

The first two terms in eq. 2.6 are the standard free energy change for an ideal solution ($\gamma = 1$), and the last term corrects for non-ideal behavior.

In the following considerations, we will return to ideal solutions with activity coefficients of $\gamma = 1$. The corresponding equations for non-ideal solutions can easily be obtained by multiplication of each concentration with the activity coefficient.

In equilibrium, ΔG becomes zero (Section 1.6), and we can write

$$\Delta G = \Delta G^0 + RT\ln\left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right)_{eq} = 0 \qquad \text{eq. 2.7}$$

The index "eq" in eq. 2.7 indicates concentrations at equilibrium. We now define the *equilibrium constant* K_{eq} as

$$K_{eq} = \left(\frac{[C]^c[D]^d}{[A]^a[B]^b}\right)_{eq} \qquad \text{eq. 2.8}$$

Note that K_{eq} is always a dimensionless number as every concentration term is a concentration relative to the concentration in the standard state! We will see in Section 2.2 how we can convert this

true equilibrium constant into a constant with dimensions related to concentrations. By combining eq. 2.7 and eq. 2.8, we then obtain

$$\Delta G^0 = -RT \ln K_{eq} \qquad \text{eq. 2.9}$$

or

$$K_{eq} = \exp\left(-\frac{\Delta G^0}{RT}\right) \qquad \text{eq. 2.10}$$

Thus, we can calculate the free energy change for the underlying reaction under standard conditions from the concentrations of the reaction partners under equilibrium conditions (relative to their respective standard states), and *vice versa*! It is important to note that the equilibrium concentrations are linked to the change in free energy under standard conditions (Box 2.1). The actual ΔG in equilibrium is always zero.

BOX 2.1: THERMODYNAMIC INFORMATION FROM EQUILIBRIUM CONSTANTS: QUANTIFYING THE ENERGETIC CONTRIBUTION OF AN AMINO ACID SIDE CHAIN TO LIGAND BINDING.

From eq. 2.9, it is evident that we can calculate free energies associated with equilibria. Specific binding of a ligand to a protein is often mediated by interactions between amino acid side chains and the ligands. We might suspect a certain amino acid, *e.g.* a tyrosine, to contribute to ligand binding by forming a hydrogen bond. We can determine the dissociation equilibrium constant of the protein-ligand complex (see Section 21.5.5.1), and can calculate the ΔG^0 of dissociation (and hence also the ΔG^0 of binding) from the equilibrium constant. Then we create a variant in which we replace the tyrosine by phenylalanine, and determine the dissociation constant of the protein-ligand complex for this variant, preferentially by the same method. If the tyrosine indeed engages in a hydrogen bond with the ligand, then removal of the hydroxyl group by the exchange to phenylalanine will lead to a decrease in ΔG^0. From the difference of the two ΔG^0 values, we can calculate a $\Delta\Delta G^0$ value that reflects the thermodynamic contribution of the tyrosine to ligand binding (under standard conditions). If the tyrosine does not contribute to ligand binding, ΔG^0 of the variant will be identical to the first value we have determined (provided that the amino acid exchange has not led to other side effects such as a conformational change of the protein).

2.1.1 Temperature dependence of the equilibrium constant

Eq. 2.10 shows that the equilibrium constant K_{eq} is temperature-dependent. Expression of ΔG^0 in eq. 2.9 by the Gibbs-Helmholtz-equation (eq. 1.111) and division by $-RT$ yields

$$\ln K_{eq} = -\frac{\Delta H^o}{RT} + \frac{\Delta S^0}{R} \qquad \text{eq. 2.11}$$

This equation is called the *van't Hoff equation*. If we plot $\ln K_{eq}$ as a function of the inverse temperature $1/T$, we thus obtain a linear plot, the *van't Hoff plot*, with the slope related to ΔH^0, and the intercept related to ΔS^0 (Figure 2.1). The plot is only linear if ΔH and ΔS are independent of temperature over the temperature range considered, *i.e.* for $\Delta C_p = 0$. If ΔC_p does not equal zero, van't Hoff plots are curved, the same way as protein stability curves are curved if $\Delta C_p \neq 0$ (see Section 1.6.2, Figure 1.26;

Section 18.5.4, Figure 18.58). To determine thermodynamic data from curved van't Hoff plots, eq. 2.11 has to be derived taking the temperature dependence of ΔH and ΔS into account. The resulting equation can then be used to describe the non-linear van't Hoff plot and to extract ΔC_p, ΔH^0, and ΔS^0. It is important to note that, particularly for biochemical reactions, the temperature range in which the equilibrium constant is determined is typically rather narrow, and the experimental data points in the van't Hoff plot are clustered. Nevertheless, the slope of the line through these points is typically well determined, and ΔH^0 can be determined reliably. In contrast, determinations of the y-axis intercept and of ΔS^0 require extrapolation from the cluster of data points over a large distance to the y-axis (Figure 2.1). Small uncertainties in the slope will therefore lead to large errors in the y-axis intercept and in ΔS^0.

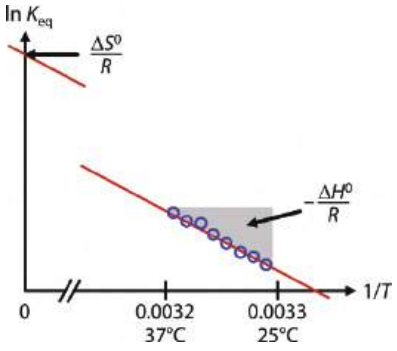


FIGURE 2.1: The van't Hoff plot. The van't Hoff equation describes the temperature dependence of equilibrium constants. A plot of lnK_{eq} as a function of the inverse temperature $1/T$ is linear if ΔH and ΔS are temperature-independent over the temperature range considered, *i.e.* if ΔC_p^0 is zero. ΔH^0 can be determined from the slope, and ΔS^0 from the y-axis intercept. The intercept has to be determined by extrapolation over a substantial inverse temperature range, leading to large errors in ΔS^0. ΔC_p^0 values $\neq 0$ lead to curved van't Hoff plots.

The van't Hoff equation is sometimes written in its differential form as

$$\frac{\mathrm{d}\ln K_{eq}}{\mathrm{d}(1/T)} = -\frac{\Delta H^o}{R} \qquad \text{eq. 2.12}$$

or

$$\frac{\mathrm{d}\ln K_{eq}}{\mathrm{d}T} = \frac{\Delta H^o}{RT^2} \qquad \text{eq. 2.13}$$

Integration from T_1 to T_2 gives the relationship

$$\ln\left(\frac{K_{eq}(T_2)}{K_{eq}(T_1)}\right) = -\frac{\Delta H^o}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right) \qquad \text{eq. 2.14}$$

which allows to calculate the equilibrium constant at the temperature T_2 if the value at T_1 and the associated change in enthalpy ΔH^0 are known.

2.1.2 The principle of Le Chatelier

In Section 1.6.2, we have discussed the pressure- and temperature dependence of G for closed systems, and have derived the equation

$$\mathrm{d}G = V\mathrm{d}p - S\mathrm{d}T \qquad \text{eq. 1.119}$$

For open systems and chemical reactions, we now have to expand this equation by the changes in chemical potential for the components to

$$dG = Vdp - SdT + \sum_i \mu_i \, dn_i \qquad \text{eq. 2.15}$$

We can infer from eq. 2.15 that G (and ΔG) depend on pressure, temperature, and amount of substance, and we have previously derived the pressure dependence of G for n,T = const., the temperature dependence for p,n = const., and the dependence on amount of substance or concentration for p,T = const. (Sections 1.6.2, 1.7.1). When equilibria are perturbed by changing one of these parameters, they will respond to this perturbation by favoring the reaction that counteracts the perturbation, and will reach a new equilibrium position. For example, an increase in temperature by transferring heat leads to a shift of the equilibrium in the endothermic direction of the chemical reaction, taking up the heat. Conversely, a decrease in temperature favors the exothermic direction, producing more heat. Likewise, an increase in pressure favors the direction of the reaction that involves a decrease in volume, and the equilibrium is shifted toward the side of the compounds with lower volume (fewer molecules, less gases), and *vice versa*. Changing the amount of substance by withdrawing a compound from the reaction favors the direction of the reaction in which this compound is produced, whereas adding the same compound in large excess favors the reaction in which it is consumed. This principle of "evading" changes in the reaction conditions is called the *principle of Le Chatelier*.

The principle of Le Chatelier is a recurrent theme in biochemical reactions. Removal of a product by a subsequent (irreversible) reaction draws an equilibrium to the side of products. The equilibrium between dihydroxyacetone phosphate and glyceraldehyde-3-phosphate, for example, is on the side of dihydroxyacetone phosphate ($K < 1$). However, in the glycolysis pathway, glyceraldehyde-3-phosphate is further metabolized *via* 1,3-bisphosphoglycerate, 3-phosphoglycerate, 2-phosphoglycerate, and phosphoenol pyruvate to pyruvate. The phosphate transfer from phosphoenol pyruvate to ADP in the last reaction is effectively irreversible, leading to a depletion of phosphoenol pyruvate. Following the principle of Le Chatelier, this removal of phosphoenol pyruvate from the equilibrium prior to the irreversible step, between 2-phosphoglycerate and phosphoenol pyruvate, causes the production of more phosphoenol pyruvate, which, in turn, leads to a depletion of 2-phosphoglycerate, and so on. Effectively, the irreversible step at the end affects all equilibria before that, eventually leading to further isomerization of dihydroxyacetone phosphate to glyceraldehyde-3-phosphate.

2.2 BINDING AND DISSOCIATION EQUILIBRIA AND AFFINITY

We have seen before that ΔG for a reverse reaction is of the same magnitude as ΔG for the forward reaction, but of opposite sign (eq. 1.114). From eq. 2.8, it follows for the associated equilibrium constants that

$$K_{\text{forward}} = \frac{1}{K_{\text{reverse}}} \qquad \text{eq. 2.16}$$

For a simple binding equilibrium of an enzyme E and its substrate S

$$\mathrm{E} + \mathrm{S} \rightleftharpoons \mathrm{ES} \qquad \text{scheme 2.2}$$

we can write the equilibrium constant $K_{\text{association}}$ for the association reaction of E and S to ES

$$K_{\text{association}} = \frac{[ES]}{[E][S]} \qquad \text{eq. 2.17}$$

following the definition of concentration of products (ES) divided by the concentration of reactants (E, S). A high association constant is equivalent to a high concentration of ES complex in equilibrium,

and we say the enzyme E has a high affinity for the substrate S. A low association constant is equivalent to a low affinity.

For the reverse reaction, the dissociation of ES to E and S, the equilibrium constant $K_{\text{dissociation}}$ is

$$K_{\text{dissociation}} = \frac{[E][S]}{[ES]} \qquad \text{eq. 2.18}$$

which is the inverse of the association constant, in accordance with eq. 2.16. A high value for the dissociation constant means a high concentration of E and S in equilibrium, and corresponds to a low substrate affinity. A low dissociation constant, in contrast, corresponds to a low concentration of free enzyme and substrate in equilibrium and a high affinity.

As discussed before (Section 2.1), all concentrations in eq. 2.17 and eq. 2.18 are relative to the concentration in the standard state, and thus dimensionless numbers. K_{eq} is therefore also a dimensionless number. For convenience, the dimensionless equilibrium constant is generally converted into a constant with units related to concentrations (Box 2.2). The unit of the association constant depends on the numbers of molecules involved in the equilibrium. For the binary binding event (Scheme 2.2), the association constant has the unit M^{-1} and the dissociation constant has the unit M. These are precisely the units we would obtain if we just entered the absolute concentrations into eq. 2.17 and eq. 2.18. For biochemical reactions in solution, we can, in fact, generally calculate equilibrium constants with units from the actual concentrations in equilibrium. This is because the standard state for a solute is 1 M, and division by 1 M does not alter the numerical value. From now on, we will always calculate equilibrium constants with dimensions for reactions in solution from absolute concentrations. The calculated value then automatically has a meaningful dimension (Box 2.2). Why is the unit meaningful? The unit M makes the dissociation constant a practical entity for argumentation and for comparison with actual concentrations. At a large excess of the substrate over enzyme ($S_0 \gg E_0$, where the index zero indicates total concentrations), we can estimate the degree of saturation of enzyme with substrate directly from the substrate concentration. At a substrate concentration that equals the K_d value, the saturation of enzyme with substrate is 50%. At much lower concentrations "below the K_d", the enzyme occurs mostly in the free form. At higher concentrations "above the K_d", the enzyme is mostly encountered as the enzyme-substrate complex. Despite the inverse relationship of the dissociation constant and the binding strength or affinity, $K_{\text{dissociation}}$ or short K_d is therefore typically used to characterize binding equilibria.

BOX 2.2: FROM A DIMENSIONLESS EQUILIBRIUM CONSTANT TO AN EQUILIBRIUM CONSTANT WITH UNITS OF CONCENTRATION.

The definition of equilibrium constants according to eq. 2.8 and the association and dissociation constants defined in eq. 2.17 and eq. 2.18 are calculated from concentrations relative to the standard state (Section 1.6.3), and therefore lead to dimensionless numbers. The dimensionless equilibrium "number" can be converted to an equilibrium constant with dimensions of (inverse) concentration, depending on the particular equilibrium. For a general reaction according to Scheme 2.1

$$aA + bB \rightleftharpoons cC + dD \qquad \text{scheme 2.1}$$

the dimension of the equilibrium constant K is

$$M^{(-a-b+c+d)} \qquad \text{eq. 2.19}$$

(*Continued*)

This is precisely the unit we would obtain if we just entered all absolute concentrations with the unit M into the definition of the equilibrium constant (eq. 2.8). It is general practice in biochemistry to calculate equilibrium constants with dimensions from absolute concentrations. This is possible because the standard state of each compound for equilibria in solution is 1 M, and the division by c^0 does not change the numerical value of the relative concentration. Nevertheless, it is important to be aware of this shortcut: often, K_d values are listed with the units µM, nM, or even pM. All these values are related to the change in free energy of the reaction under standard conditions, ΔG^0. To calculate ΔG^0 from the equilibrium constant, the value in M has to be used, otherwise the conversion is incorrect. Using concentrations relative to the standard state avoids this complication. Also, the use of equilibrium constants with units can be confusing when water is not only the solvent but also a reactant. For the solvent, the standard concentration is that of the pure solvent. As the solvent by definition is in large excess, the change in concentration due to the reaction is typically very small. In other words, the actual concentration almost equals the concentration of pure water. The ratio $c(H_2O)/c^0(H_2O)$ can therefore be approximated as unity, and the concentration term for water can be ignored, which is not so obvious when absolute concentrations are used.

From the definition of the dissociation constant K_d, we can calculate the concentration of the ES complex in equilibrium as a function of total substrate and enzyme concentrations, E_0 and S_0. First, we state the conservation of mass for $[E]$ and $[S]$:

$$E_0 = [E] + [ES] \qquad \text{eq. 2.20}$$

and

$$S_0 = [S] + [ES] \qquad \text{eq. 2.21}$$

We can then express K_d as

$$K_d = \frac{(E_0 - [ES])(S_0 - [ES])}{[ES]} \qquad \text{eq. 2.22}$$

and obtain the quadratic equation

$$0 = [ES]^2 - (E_0 + S_0 + K_d) \cdot [ES] + E_0 S_0 \qquad \text{eq. 2.23}$$

This quadratic equation has two solutions (see Section 31.2)

$$[ES] = \frac{E_0 + S_0 + K_d}{2} \pm \sqrt{\left(\frac{E_0 + S_0 + K_d}{2}\right)^2 - E_0 S_0} \qquad \text{eq. 2.24}$$

of which only the solution with the negative sign in front of the square root is physically meaningful (the other solution gives $[ES] > E_0$ or $[ES] > S_0$, which is impossible). The concentration of the ES complex at any given total concentration of E_0 and S_0 as a function of the dissociation constant K_d is

$$[ES] = \frac{E_0 + S_0 + K_d}{2} - \sqrt{\left(\frac{E_0 + S_0 + K_d}{2}\right)^2 - E_0 S_0} \qquad \text{eq. 2.25}$$

By dividing eq. 2.25 by the total enzyme concentration E_0, we can express the average number of binding sites occupied as a function of the substrate concentration, $\bar{\nu}$, also termed the degree of binding:

$$\bar{\nu} = \frac{[ES]}{E_0} = \frac{\frac{E_0 + S_0 + K_d}{2} - \sqrt{\left(\frac{E_0 + S_0 + K_d}{2}\right)^2 - E_0 S_0}}{E_0} \qquad \text{eq. 2.26}$$

The corresponding equations for multiple binding sites, both interacting and non-interacting, will be derived in Section 3.4. We will return to eq. 2.25 in the context of binding kinetics (Chapter 11). Eq. 2.25 also becomes important when we discuss how to determine K_d values experimentally using spectroscopic probes (Section 21.5.5.1).

2.3 PROTOLYSIS EQUILIBRIA: THE DISSOCIATION OF ACIDS AND BASES IN WATER

A specific binding/dissociation event with important implications on biochemical reactions is the dissociation of an acid HA or a base BOH in water. Acidic (carboxyl, phosphoryl, sulfonyl, thiol) and basic (amine, imidazole, and guanidine) groups occur in numerous biomolecules, including metabolites (see Box 18.4), fatty acids and phospholipids (see Section 18.4), amino acids as the building blocks of proteins (see Section 18.1), and in both nucleobases and nucleic acids (see Section 19.1). The presence of such groups is responsible for the buffering effect of plasma and the pH dependence of reactions of these molecules (see Section 13.3).

Protolysis equilibria are described by the equilibrium constants K_A and K_B:

$$K_A = \frac{[A^-][H^+]}{[HA]} \qquad \text{eq. 2.27}$$

for the dissociation of the acid HA and

$$K_B = \frac{[B^+][OH^-]}{[BOH]} \qquad \text{eq. 2.28}$$

for the dissociation of the base BOH in water. Often, the negative decadic logarithms of the K_A and K_B values are tabulated (see also Table 18.1 and Table 30.1). In analogy to the pH value, these values are called pK_A and pK_B. The pK_A and pK_B values are very convenient constants (see eq. 2.33): at a pH that equals the pK_A, half of the acid is deprotonated, the other half is protonated. At pH values well above the pK_A, the acid is mostly dissociated, and at pH values well below the pK_A, it is mostly protonated. Correspondingly, half of the base is dissociated at a pH that equals the pK_B, the other half is not. At pH values above the pK_B, the dissociation is less, and at pH values below the pK_B, the base is more and more dissociated.

The dissociated form of an acid HA, the anion A^-, can accept a proton from water and act as a base. Similarly, the pronated form of a base B, BH^+, can transfer a proton to water and act as an acid. Such pairs of HA and A^- are called acid and *conjugate base*. The pK_A and pK_B values for acids and their conjugate bases are related: the K_B for the reaction of A^- with H_2O is

$$K_B = \frac{[HA][OH^-]}{[A^-]} \qquad \text{eq. 2.29}$$

By multiplying K_A (eq. 2.27) and K_B, we obtain

$$K_A K_B = \frac{[A^-][H^+]}{[HA]} \frac{[HA][OH^-]}{[A^-]} = [H^+][OH^-] = K_W \qquad \text{eq. 2.30}$$

The product of K_A and K_B of a conjugate acid and base thus equals the ionic product K_W of water. The ionic product is a constant, with $K_W = 10^{-14}$. The small value indicates that the equilibrium of water dissociation into H^+ and OH^- lies very much on the side of undissociated water. By taking the negative decadic logarithm of eq. 2.30, we obtain

$$pK_A + pK_B = -\log K_W = 14 \qquad \text{eq. 2.31}$$

Hence, we can directly calculate the pK_B for a conjugate base if we know the pK_A for dissociation of the acid.

Taking the negative decadic logarithm of eq. 2.27 gives

$$-\log K_A = -\log \frac{[A^-][H^+]}{[HA]} = -\log \frac{[A^-]}{[HA]} - \log[H^+] \qquad \text{eq. 2.32}$$

or

$$\mathrm{pH} = pK_A + \log \frac{[A^-]}{[HA]} \qquad \text{eq. 2.33}$$

Eq. 2.33 is the *Henderson-Hasselbalch equation* that describes the pH of a solution as a function of the concentration ratio of acid and conjugate base. For $[A^-] = [HA]$, $\mathrm{pH} = pK_A$. Under these conditions, the solution of acid and conjugate base acts as a pH buffer: addition of acid or base leads to very small changes in pH. The *buffer capacity* β is defined as the amount of acid (or base) added, divided by the change in pH:

$$\beta = \frac{[H^+]_{\text{added}}}{\Delta \mathrm{pH}} \qquad \text{eq. 2.34}$$

Many acids, such as oxaloacetic acid, phosphoric acid, or amino acids, undergo multiple protolysis steps with different pK_A values, and buffer the pH of a solution at different pH values.

2.4 THERMODYNAMIC CYCLES, LINKED FUNCTIONS, AND APPARENT EQUILIBRIUM CONSTANTS

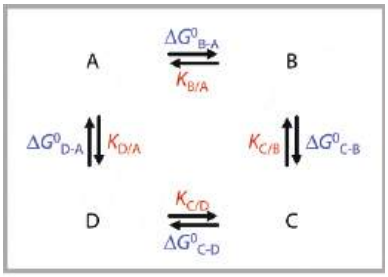
In biochemistry, equilibria are usually not isolated, but linked by compounds that are involved in several equilibria. Often, one species A can be converted into a different species C by alternative pathways, *e.g.* either *via* B or D (Figure 2.2). Such a pattern is called a *thermodynamic cycle*. Similar to a Carnot cycle (Section 1.3.5.1), we can start from A, convert it to B and C and D and back to A. The associated change in free energy ΔG is zero because initial and final states are identical and have the same free energy G. Similarly, the free energy change from A through B to C has to be identical to the free energy change for conversion of A through D to C because initial and final states are identical, and the pathway we chose to move from A to C is irrelevant for the change in the state function G. Under standard conditions, we can therefore write

$$\Delta G^0_{B-A} + \Delta G^0_{C-B} = \Delta G^0_{D-A} + \Delta G^0_{C-D} \qquad \text{eq. 2.35}$$

Note that $\Delta G^0{}_{Y-X}$ corresponds to the ΔG^0 value associated with the reaction from X to Y. In terms of the individual equilibrium constants, this corresponds to

$$K_{B/A} \cdot K_{C/B} = K_{D/A} \cdot K_{C/D} \qquad \text{eq. 2.36}$$

$K_{Y/X}$ is the equilibrium constant for the equilibrium between X and Y. Eq. 2.35 and eq. 2.36 show that the four equilibria are *thermodynamically linked*. If we know ΔG^0 or K for three of the four equilibria, we can therefore directly calculate the missing fourth value.



$$\Delta G^0_{B-A} + \Delta G^0_{C-B} = \Delta G^0_{D-A} + \Delta G^0_{C-D}$$

$$K_{B/A} \cdot K_{C/B} = K_{D/A} \cdot K_{C/D}$$

FIGURE 2.2: A thermodynamic cycle. Compound A can be converted into compound C following two pathways, A→B→C or A→D→C. The four equilibria form a thermodynamic cycle. If we move through all four reactions from A to B to C to D and back to A, the initial and final states are identical, and the change in free energy must be zero. An equivalent statement to formulate the thermodynamic balance is that the free energy change for both pathways from A to C must be identical, *i.e.* $\Delta G^0_{B/A} + \Delta G^0_{C/B} = \Delta G^0_{D/A} + \Delta G^0_{C/D}$. As a consequence, the four equilibrium constants are also linked: $K_{B/A} \cdot K_{C/B} = K_{D/A} \cdot K_{C/D}$. Note that here all equilibrium constants are depicted in the center of the thermodynamic cycle, whereas all free enthalpy values are depicted on the outside. For each individual equilibrium, the equilibrium constant and the ΔG^0 value both are defined for the same direction of the reaction. See text for definitions of the equilibrium constants.

From eq. 2.36, we can also conclude that the ratios of the equilibrium constants for the two vertical processes (A→D, B→C) and for the two horizontal processes (A→B, D→C) must be equal:

$$\frac{K_{D/A}}{K_{C/B}} = \frac{K_{B/A}}{K_{C/D}} \qquad \text{eq. 2.37}$$

This insight is useful if we look at a common example of a biochemical thermodynamic cycle, the binding of two substrates to an enzyme (Figure 2.3). The enzyme can either bind substrate A first, and the EA complex then binds substrate B, or substrate B binds first, and the EB complex then binds substrate A. In both cases, the final product is the EAB complex. The binding equilibria form a thermodynamic cycle, and we know that

$$K_A \cdot K'_B = K'_A \cdot K_B \qquad \text{eq. 2.38}$$

where K_A is the equilibrium constant for the dissociation of the EA complex, K_A' the dissociation constant of A from EAB, K_B the dissociation constant of the EB complex, and K_B' the dissociation constant for B from the EAB complex. Binding of A and B can be independent, and in this case, K_A equals K_A', and K_B equals K_B'. It is also possible that binding of A promotes binding of B. In this case, B will be bound more tightly by the EA complex than by E alone. In terms of equilibrium dissociation constants, this means K_B' is decreased compared to K_B (Section 2.2; eq. 2.18). From eq. 2.38, it follows that in this case B must also favor binding of A by the same factor α, *i.e.*

$$\alpha = \frac{K_A}{K'_A} = \frac{K_B}{K'_B} \qquad \text{eq. 2.39}$$

Thus, if binding of A favors binding of B, then for thermodynamic reasons binding of B must favor binding of A to the same extent! Similarly, if binding of A disfavors binding of B ($K_B' > K_B$), then binding of B must, in turn, disfavor binding of A by the same factor. This mutual effect because of thermodynamically coupled equilibria is referred to as *thermodynamic linkage*.

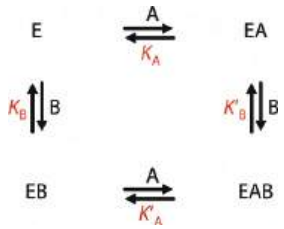


FIGURE 2.3: Thermodynamic linkage in substrate binding. An enzyme E binds two substrates, A and B, to form the EAB complex. K_A is the dissociation constant of the EA complex, K_B the dissociation constant of the EB complex. K_A' and K_B' are the corresponding constants in the presence of the second ligand. The four equilibria form a thermodynamic cycle, and the equilibrium constants are linked. If binding of A favors binding of B ($K_B' < K_B$), then binding of B also favors the binding of A to the same extent ($K_A' < K_A$) due to the thermodynamic balance within the cycle.

If the binding of A and B is interdependent, we can calculate the *interaction energy* $\Delta\Delta G^0_{int}$ for A and B binding by comparing the free energy changes upon formation of the EAB complex when binding of A and B is independent with the free energy change when binding is coupled. Note that the ΔG^0 values in eq. 2.35 have been calculated from the equilibrium constants of dissociation, and are changes in free energy upon dissociation. The ΔG^0 for dissociation of EAB in the absence of coupling is $\Delta G^0_A + \Delta G^0_B$, for coupling it is $\Delta G^0_A + \Delta G^{0\prime}_B$ for EAB→EA→E or $\Delta G^{0\prime}_A + \Delta G^0_B$ for EAB→EB→E. The interaction energy $\Delta\Delta G^0_{int}$ is the difference between these values:

$$\Delta\Delta G^0_{int} = \Delta G^0_A + \Delta G^0_B - \left(\Delta G^0_A + \Delta G^{0\prime}_B\right) = \Delta G^0_A + \Delta G^0_B - \left(\Delta G^{0\prime}_A + \Delta G^0_B\right) \qquad \text{eq. 2.40}$$

$\Delta\Delta G^0_{int}$ can be positive or negative. $\Delta\Delta G^0_{int} < 0$ means binding of one substrate increases the affinity for the second substrate, we say that binding of A and B shows *positive thermodynamic linkage*. In cases where $\Delta\Delta G^0_{int} > 0$, binding of one substrate decreases the affinity of the enzyme for the second substrate, the two substrates bind with *negative thermodynamic linkage*. Sometimes this phenomenon is also referred to as positive or negative cooperativity, although the term *cooperativity* describes a different effect (Section 13.5). Note that we can also calculate $\Delta\Delta G^0_{int}$ from the free energies of binding. In this case, $\Delta\Delta G^0_{int}$ is positive for positive thermodynamic linkage, and negative for negative thermodynamic linkage. Interaction energies can be determined in double mutant cycles (Box 2.3).

BOX 2.3: INTERACTION ENERGY: DOUBLE MUTANT CYCLES.

Thermodynamic cycles can be used to identify interactions between amino acids in a protein. In such an approach, the two amino acids of interest are exchanged to different amino acids that cannot engage in this interaction, most commonly alanine. The two amino acids are replaced individually and in combination. For all three modified (variant) proteins and the unmodified (wild-type) protein, the thermodynamic stability is determined experimentally (Section 18.5.4). The contribution of the amino acid to the stability is determined as the difference in ΔG^0_{unfold} for the variant compared to the wild-type ($\Delta\Delta G^0 = \Delta G^0_{unfold,wild-type} - \Delta G^0_{unfold,variant}$). If the two amino acids, *e.g.* a serine and a tyrosine, engage in an interaction that contributes to the thermodynamic stability of the protein, then exchanging either of them (*e.g.* serine to alanine or tyrosine to phenylalanine, Figure 2.4) will reduce the stability of the protein by this

(*Continued*)

contribution. For variant 1 we obtain a destabilization by $\Delta\Delta G^0_1$, for variant 2 by $\Delta\Delta G^0_2$. If the two amino acids do not interact with each other, but with other partners, $\Delta\Delta G^0_1$ and $\Delta\Delta G^0_2$ are likely different. For the double-variant, we will obtain a $\Delta\Delta G^0_3$ that is the sum of $\Delta\Delta G^0_1$ and $\Delta\Delta G^0_2$ because we now remove both interactions. If the two amino acids interact with each other, however, then $\Delta\Delta G^0_1$ and $\Delta\Delta G^0_2$ are the same because the same interaction is removed. Substitution of the second amino acid in the double-variant will not have any further destabilizing effect, and $\Delta\Delta G^0_3$ will equal $\Delta\Delta G^0_1$ and $\Delta\Delta G^0_2$. A similar approach can be used to identify cooperative interactions of two side chains in substrate binding, for example. Here, $\Delta\Delta G^0$ is the difference in free energy change of dissociation of the enzyme-substrate complex between wild-type and variant ($\Delta\Delta G^0 = \Delta G^0_{diss,wild-type} - \Delta G^0_{diss,variant}$). In this case, the energies can be additive ($\Delta\Delta G^0_3 = \Delta\Delta G^0_1 + \Delta\Delta G^0_2$; independent interactions of the two side chains with the substrates) or non-additive ($\Delta\Delta G^0_3 \neq \Delta\Delta G^0_1 + \Delta\Delta G^0_2$; cooperative interactions with the substrate). For $\Delta\Delta G^0_3 > (\Delta\Delta G^0_1 + \Delta\Delta G^0_2)$ the two amino acids show positive cooperativity in substrate binding, for $\Delta\Delta G^0_3 < (\Delta\Delta G^0_1 + \Delta\Delta G^0_2)$ they show negative cooperativity. The difference $(\Delta\Delta G^0_1 + \Delta\Delta G^0_2) - \Delta\Delta G^0_3$ is what we have defined as the interaction energy $\Delta\Delta G^0_{int}$ in a thermodynamic cycle (see text). If we calculate $\Delta\Delta G^0_{int}$ from dissociation free energies, a positive value corresponds to negative, a negative value to positive thermodynamic linkage. For binding free energies, a positive or negative $\Delta\Delta G^0_{int}$ corresponds to positive or negative thermodynamic linkage.

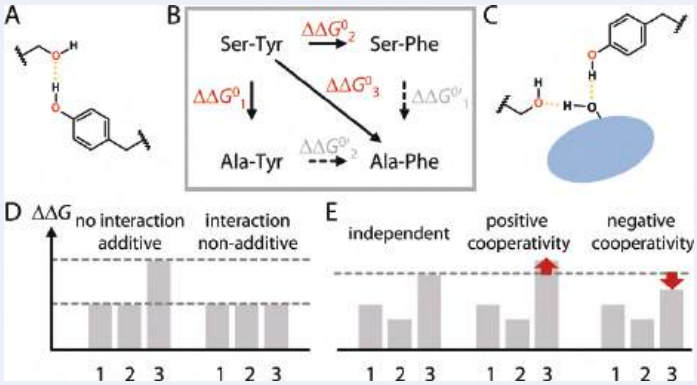


FIGURE 2.4: Double mutant cycles to identify interactions between side chains or cooperativity in substrate binding. Interacting side chains (A) or cooperativity of two side chains in binding of substrate (blue; C) can be identified in double mutant cycles (B). In a double mutant cycle, the amino acids of interest (here Ser and Tyr) are individually and simultaneously replaced by other side chains that cannot engage in the interaction. The wild-type protein and the three variants form a thermodynamic cycle. The $\Delta\Delta G^0$ values are either differences in thermodynamic stability between wild-type and protein (bottom left) or differences in substrate binding energies (bottom right). $\Delta\Delta G^0_1$, $\Delta\Delta G^0_2$, and $\Delta\Delta G^0_3$ are measured experimentally. D: If the two side chains interact with each other, this interaction stabilizes the protein. Removal of the interaction by either side chain exchange leads to a decrease in thermodynamic stability by $\Delta\Delta G^0_1$ or $\Delta\Delta G^0_2$, respectively. Exchanging the second side chain does not have an additional destabilizing effect ($\Delta\Delta G^0_1 = \Delta\Delta G^0_2 = \Delta\Delta G^0_3$). If the two side chains interact with different partners, however, exchange of both side chains will remove both of these interactions, and $\Delta\Delta G^0_3$ is the sum of $\Delta\Delta G^0_1$ and $\Delta\Delta G^0_2$. E: If two side chains interact with a ligand, exchange of either of these amino acids leads to reduced binding energy ($\Delta\Delta G^0_1$, $\Delta\Delta G^0_2$). If the interactions are independent, simultaneous removal of both interactions leads to a change in binding energy that is the sum of the individual contributions (left). The $\Delta\Delta G$ values are not additive if there is positive (right) or negative cooperativity (center) between these interactions in substrate binding. The arrows indicate the interaction energy (see text).

A second important example of a thermodynamic cycle is represented by the following situation: an enzyme exists in two different conformations, E and E'. The enzyme contains a basic side chain that can be protonated to EH^+ or $E'H^+$ (Figure 2.5). The equilibrium constants for the conformational equilibrium are K (deprotonated state) and K' (protonated state), the equilibrium constants for the protonation equilibrium are K_A (E) and K'_A (E'). The four species E, E', EH^+, and $E'H^+$ are linked by four equilibria that form a thermodynamic cycle (Figure 2.5).

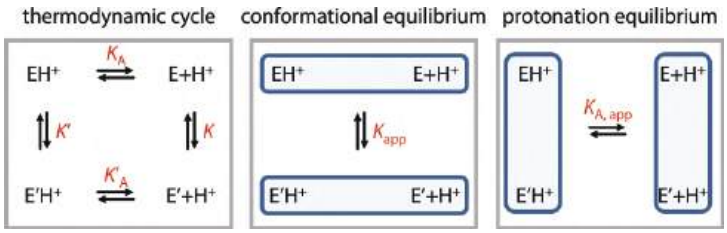


FIGURE 2.5: Thermodynamic cycles and linked functions. The thermodynamic cycle (left) shows coupled equilibria of protonation and conformational equilibria of an enzyme E. The enzyme can exist in two states E and E′. Both E and E′ can be protonated. The equilibrium constants for the individual equilibria are indicated. We can now focus on the overall conformational equilibrium by summing all species of the enzyme in conformational state E (E and EH^+) as one pseudo-species, and all species in state E′ (E′ and $E'H^+$) as a second pseudo-species, and defining an apparent equilibrium constant between these species (center). The apparent constant for the conformational equilibrium will then depend on the pH. Alternatively, we can focus on the protonation equilibrium, and sum together all deprotonated species (E and E′) and all protonated species (EH^+ and $E'H^+$) as pseudo-species, and define an apparent equilibrium constant between those (right). The apparent protolysis constant will depend on the conformational equilibria: protonation and conformational change are linked functions.

If K_A and K_A' are identical, then K and K' are also identical: the conformational equilibrium is independent of the protonation state of the enzyme. Often, the two types of equilibria are coupled, and $K_A \neq K_A'$ and $K \neq K'$. In this case, we can define an *apparent equilibrium constant* for the overall conformational equilibrium, K_{app}, by collecting all species of E in the denominator, and all species of E′ in the numerator (Figure 2.5). The apparent equilibrium constant is then

$$K_{app} = \frac{[E'] + [E'H^+]}{[E] + [EH^+]} \qquad \text{eq. 2.41}$$

The protonation equilibria of E and E′ are described by the equilibrium constants K_A and K_A'. Protonation equilibria are typically formulated as dissociation reactions (protolysis; Section 2.3).

$$K_A = \frac{[E][H^+]}{[EH^+]} \qquad \text{eq. 2.42}$$

and

$$K'_A = \frac{[E'][H^+]}{[E'H^+]} \qquad \text{eq. 2.43}$$

We can now solve eq. 2.42 and eq. 2.43 for EH^+ and $E'H^+$, respectively, and substitute these expressions into eq. 2.41 to obtain

$$K_{app} = \frac{[E'] + \frac{[E'][H^+]}{K'_A}}{[E] + \frac{[E][H^+]}{K_A}} \qquad \text{eq. 2.44}$$

By factoring out $[E']$ in the numerator and $[E]$ in the denominator, eq. 2.44 simplifies to

$$K_{app} = \frac{[E']\left(1 + \frac{[H^+]}{K'_A}\right)}{[E]\left(1 + \frac{[H^+]}{K_A}\right)} \qquad \text{eq. 2.45}$$

Substitution of $[E']/[E]$ by K gives

$$K_{app} = K\frac{\left(1 + \frac{[H^+]}{K'_A}\right)}{\left(1 + \frac{[H^+]}{K_A}\right)} \qquad \text{eq. 2.46}$$

We see that as a result of the coupling of the conformational and protonation equilibria, the apparent equilibrium constant for the conformational equilibrium is dependent on the H^+ concentration and thus on the pH. We can distinguish two limiting cases: at high pH, the proton concentration is small, and the terms H^+/K_A and H^+/K'_A in the sums in the numerator and denominator of eq. 2.46 are small and can be neglected, giving

$$K_{app,high\ pH} = K \qquad \text{eq. 2.47}$$

Under these conditions, EH^+ or $E'H^+$ will not be populated. The conformational equilibrium will be established between E and E′, and K_{app} equals K. At low pH, *i.e.* high proton concentrations, the terms H^+/K_A and H^+/K'_A in eq. 2.46 are large and now dominate the sums in numerator and denominator. Eq. 2.46 then simplifies to

$$K_{app,low\ pH} = K\frac{K_A}{K'_A} \qquad \text{eq. 2.48}$$

Note that, according to the thermodynamic balance in the cycle, the right-hand side of eq. 2.48 equals K'! At low pH, both E and E′ will be protonated, and the conformational equilibrium will be dominated by the equilibrium between EH^+ and $E'H^+$. The behavior of K_{app} as a function of the pH between these two limiting cases is illustrated in Figure 2.6. The conformational equilibrium becomes pH-dependent because the protonation equilibria are different for the two conformations.

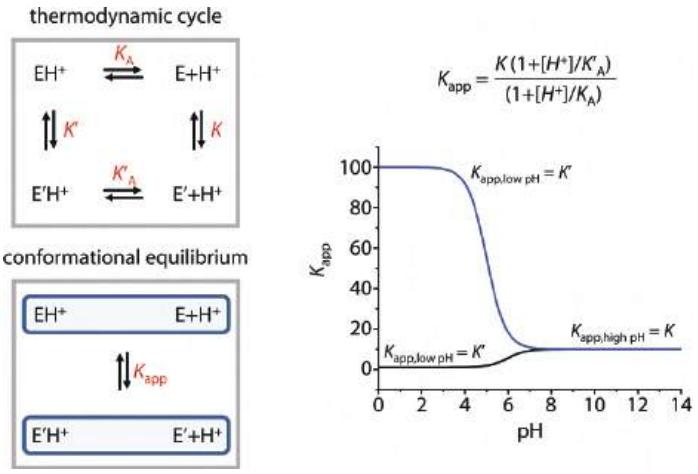


FIGURE 2.6: Linked functions and apparent equilibrium constants. K_{app} as a function of pH. The apparent equilibrium constant for the conformational equilibrium, K_{app}, depends on the pH. K_A and K_A' are the protolysis constants for the enzyme in the two conformations. Black curve: K_A = 1 μM (pK_A = 6), K_A' = 10 μM (pK_A' = 5), K = 10; blue curve: K_A = 10 μM (pK_A = 5), K_A' = 1 μM (pK_A' = 6), K = 10. At low pH, the enzyme is protonated, and K_{app} approaches K'. At high pH, the enzyme is deprotonated, and K_{app} approaches K. The inflection point of the pH dependence is at $[H^+] = K_A$ or pH = pK_A.

We can also define an apparent protonation constant by collecting all protonated and deprotonated species (Figure 2.5). By convention, we formulate the equilibrium constant $K_{A,app}$ as a dissociation constant:

$$K_{A,app} = \frac{\left([E'] + [E]\right)[H^+]}{[E'H^+] + [EH^+]} \qquad \text{eq. 2.49}$$

We can now substitute the concentrations for E′ by $K[E]$ and E′H$^+$ by $K'[EH^+]$ and obtain

$$K_{A,app} = \frac{\left(K[E] + [E]\right)[H^+]}{K'[EH^+] + [EH^+]} \qquad \text{eq. 2.50}$$

Now $[E]$ and $[EH^+]$ can be factored out, giving

$$K_{A,app} = \frac{(K+1)[E][H^+]}{(K'+1)[EH^+]} \qquad \text{eq. 2.51}$$

The ratio of the concentrations in eq. 2.51 is K_A, and we can simplify to

$$K_{A,app} = K_A \frac{K+1}{K'+1} \qquad \text{eq. 2.52}$$

The true protolysis constant K_A of the protonated basic group in the enzyme is modified by the factor $(K+1)/(K'+1)$ because of the coupling to the conformational equilibrium. This factor is significantly different from unity when $K \neq K'$ and either K or K' are larger than unity (Figure 2.7).

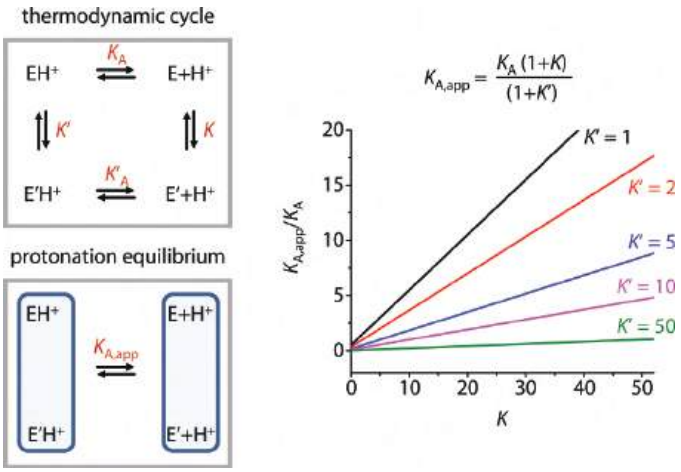


FIGURE 2.7: Linked functions and apparent equilibrium constants: $K_{A,app}/K_A$ for different values of K and K'. The apparent protolysis equilibrium constant, $K_{A,app}$, is the intrinsic protolysis constant K_A, multiplied by $(K+1)/(K'+1)$. K and K' are the equilibrium constants for the conformational equilibria of the deprotonated and protonated enzyme. The ratio of $K_{A,app}/K_A$ is shown as a function of K for several values of K'. A coupled conformational equilibrium shifts the apparent pK_A value of an ionizable group when K and K' are different, and one of them is large.

An example of such behavior is chymotrypsin. α-chymotrypsin is a serine protease that consists of three peptide chains, A, B, and C, linked by disulfide bonds. The protonation state of the N-terminus of chain B strongly influences the conformational equilibrium of α-chymotrypsin: while the enzyme is predominantly in the inactive E state when it is deprotonated ($K < 1$), it is predominantly in the active E' state in the protonated form ($K' \gg 1$). From the thermodynamic balance, we know that the K_A value of the ionizable group must also strongly differ in E and E'. According to eq. 2.52, $K_{A,app}$ is much smaller than K_A. Thus, a change in pH not only shifts the protonation equilibrium, but also dramatically alters the conformational equilibrium of chymotrypsin. At low pH, the enzyme is in the protonated state, and $E'H^+$ is the predominant form. Raising the pH leads to a dissociation of $E'H^+$ to E' and H^+. In the deprotonated state, the conformational equilibrium is on the side of E, and E' is converted to E. The overall equilibrium is established between $E'H^+$ and E (Figure 2.8).

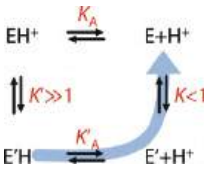


FIGURE 2.8: Linked ionization and conformational equilibria in α-chymotrypsin. α-chymotrypsin exists in an active conformation (E') in which the N-terminus of the polypeptide chain B is protonated ($E'H^+$). The conformational equilibrium between E and E' is characterized by an equilibrium constant $K < 1$ in the deprotonated state, and $K' \gg 1$ in the protonated state. At low pH (high H^+ concentration), the enzyme is protonated, and because K' is much larger than unity, $E'H^+$ is the predominant species. With increasing pH, $E'H^+$ dissociates, and because K is less than unity, E' is converted to E (blue arrow).

By deriving the apparent equilibrium constants K_{app} and $K_{A,app}$, we have seen that the conformational and protonation equilibria are coupled, or linked. Thermodynamically coupled properties are also called *linked functions*. Apparent equilibrium constants are often also called group or *macroscopic equilibrium constants*. In contrast, K_A, K'_A, K, and K' are *microscopic equilibrium constants*. Note that apparent equilibrium constants are the only case where we have sums of concentrations in the numerator and the denominator! Microscopic equilibrium constants always have products of concentrations in the numerator and denominator.

QUESTIONS

2.1 A reaction A+B ⇄ C+D leads to equilibrium concentrations of $[A]_{eq}$ = 1 µM, $[B]_{eq}$ = 2 µM, $[C]_{eq}$ = 4 µM, $[D]_{eq}$ = 8 µM. Calculate K_{eq} and ΔG^0. What is the maximum work the system can perform when 5 µM A and 10 µM B, or 1 µM C and 1 µM D are mixed? T = 298.15 K.

2.2 A reaction A+B ⇄ C has an equilibrium constant

$$K_{eq} = \frac{[C]_{eq}}{[A]_{eq}[B]_{eq}} = 5\ \mathrm{mM}^{-1}$$

What are the equilibrium concentrations of A and B when you start the reaction from 2 mM C? What are the equilibrium concentrations if you mix 1 mM A and 2 mM B?

2.3 $\Delta G^{0'}$ for ATP hydrolysis under standard conditions (pH 7, 25°C) is –31 kJ mol^{-1}. Calculate ΔG for steady-state conditions in the living cell of $[ATP] = 10^{-3}$ M, $[ADP] = 10^{-4}$ M, and $[phosphate] = 10^{-2}$ M.

2.4 A reaction A ⇄ B has a free energy change of $\Delta G^{0'}$ = 25 kJ mol^{-1} under standard conditions. The $\Delta G^{0'}$ for ATP hydrolysis is –31 kJ mol^{-1}. What is the minimum ATP/ADP ratio required to drive the reaction towards B for $[P_i]$ = 10 mM, $[A]$ = 5 mM, $[B]$ = 1 mM?

2.5 Adenylate cyclase catalyzes the formation of cyclic AMP (cAMP) from ATP:

ATP → cAMP + PP_i

The equilibrium constant for this reaction at standard conditions is K_{eq} = 0.065. $\Delta G^{0'}$ for ATP hydrolysis to AMP is –33.5 kJ mol^{-1}. What is $\Delta G^{0'}$ for the hydrolysis of cAMP to AMP?

2.6 Consider the following scheme of reactions from the citric acid cycle:

fumarate + H_2O → malate	$K_{eq,1}$ = 4.5
malate + NAD^+ → oxaloacetate + NADH + H^+	$K_{eq,2} = 1.3 \cdot 10^{-5}$
oxaloacetate + acetyl-CoA + H_2O → citrate + CoA-SH	$K_{eq,3} = 3.2 \cdot 10^{5}$

What is the overall free energy change for the conversion of fumarate to citrate?

2.7 The K_d of two enzyme-substrate complexes differs by a factor of ten: $K_{d,1} = 10 \cdot K_{d,2}$. What is the energetic difference $\Delta\Delta G$ in free energies of binding?

2.8 An enzyme binds its substrate with 1:1 stoichiometry. The equilibrium association constant is $K_{a,S} = 10^6$ M^{-1}. What is the equilibrium dissociation constant? Calculate the degree of binding, $\nu = [ES]/E_0$ at a concentration of 1 µM enzyme and 1 µM, 10 µM, and 100 µM substrate.

2.9 The equilibrium between the folded and denatured state of a protein is characterized by an equilibrium constant of folding, K_{fold}. Calculate the fraction θ of the folded protein.

2.10 Phosphoglyerate kinase couples the endergonic phosphorylation of 3-phosphoglycerate (3-PG) to 1,3-bisphosphoglycerate with the exergonic reaction of ATP hydrolysis. $\Delta G^{0'}$ for ATP hydrolysis is –31 kJ mol^{-1}, $\Delta G^{0'}$ for the phosphorylation of 3-PG to 1,3-BPG is +50 kJ mol^{-1}.
What is $\Delta G^{0'}$ for the coupled reaction? In yeast, the ATP/ADP ratio is 10 under optimal conditions, 2 under starving conditions. For what values of 3-PG/1,3-BPG does the overall reaction become exergonic?

2.11 A membrane transporter/cargo complex has a K_d value of 10 µM at 25°C, at 37°C, the K_d increases to 25 µM. What is ΔS and ΔH of binding at 25°C?

2.12 An acid HA has a dissociation constant K_A = 0.1 mM. What are the pH of a 1 M and a 0.1 mM aqueous solution of the acid, a 1 M solution of the sodium salt of the acid, and a mixture of the sodium salt and the acid, both at a concentration of 1 M? $K_w = [OH^-][H^+] = 10^{-14}$

2.13 The Tris base has a pK_A of 8.3. What is the pH change for 1 L of a 50 mM Tris buffer at pH 8.3 upon addition of 10 mmol H^+? How much would the pH change when the same amount of protons is added to 1 L of unbuffered water of pH 8.3? What is the buffer capacity β of 50 mM Tris, pH 8.3, and water at pH 8.3? How much would the buffer capacity change if the Tris buffer had a concentration of 200 mM?

2.14 The protonation enthalpy for Tris is −4.2 kJ mol^{-1}, for hydrogen phosphate (HPO_4^{2-}) it is −46 kJ mol^{-1}. What is the change in pK_A with temperature for both buffers? Which buffer would you use to investigate thermal unfolding of a protein?

2.15 A 2 mM solution of ATP is hydrolyzed at pH 8.0 in water and in 50 mM Tris buffer. What is the change in pH? pK_A(Tris) = 8.3.

2.16 The isoelectric point pI of amino acids is the pH at which the net charge is zero. Calculate the pI as a function of the pK_A values for amino acids with acidic side chains, with basic side chains, and with side chains that are neither basic nor acidic.

2.17 Calculate the fractions of all histidine species at pH 4, 6, and 8. pK_{A1} = 1.7 (carboxylic group), pK_{A2} = 6.0 (side chain), pK_{A3} = 9.1 (amino group).

2.18 An enzyme E binds a substrate S and a cofactor C. The equilibrium dissociation constant $K_{d,S}$ of the enzyme-substrate complex ES is 1 μM, for EC it is 10 μM. When the cofactor C is present, $K'_{d,S}$ is decreased to 0.1 μM. What is the value for the dissociation constant $K'_{d,C}$ of the enzyme-cofactor complex in the presence of substrate S? Calculate the interaction energy $\Delta\Delta G_{int}$ for cofactor and substrate binding.

2.19 A protein exists in two alternative conformations N_1 and N_2 in the native state and unfolds to U from N_2:

$$N_1 \overset{K_1}{\rightleftharpoons} N_2 \overset{K_2}{\rightleftharpoons} U$$

Calculate the apparent equilibrium constant for folding.

2.20 An enzyme exists in an autoinhibited (AI) and an uninhibited, active state (E). The equilibrium constant of autoinhibition, K_{ai}, is

$$K_{ai} = \frac{[AI]}{[E]}$$

Both forms of the enzyme interact with its ligand L. Formulate the dissociation constants for the two complexes, and calculate the apparent dissociation constant of the enzyme-ligand complex. In Chapter 3, the same task is solved using statistical thermodynamics considerations.

2.21 A ligand binds more tightly to the folded state (N) of a protein than to the unfolded state (U). Show that the ligand stabilizes the protein and calculate by how much ($\Delta\Delta G_{fold}$ = ?).

2.22 One effect of denaturants is a preferred interaction with the unfolded state of proteins. Show how this leads to destabilization and unfolding.

REFERENCES

Horovitz, A. (1996) Double-mutant cycles: A powerful tool for analyzing protein structure and function. *Fold Des* **1**(6): R121–126.
· review about the application of double mutant cycles to proteins

Horovitz, A. and Fersht, A. R. (1990) Strategy for analysing the co-operativity of intramolecular interactions in peptides and proteins. *J Mol Biol* **214**(3): 613–617.
· theoretical treatment of interaction energies and cooperativity of interactions

Klostermeier, D. and Millar, D. P. (2002) Energetics of hydrogen bond networks in RNA: Hydrogen bonds surrounding G+1 and U42 are the major determinants for the tertiary structure stability of the hairpin ribozyme. *Biochemistry* **41**(48): 14095–14102.
· application of double-mutant cycles to probe interactions in RNA



3 Statistical thermodynamics

Thermodynamic functions describe the average behavior of large ensembles of molecules. The behavior of these macroscopic systems is determined by the energies of their component molecules. We have discussed the molecular origin of enthalpy, entropy, and the free energy before in a qualitative manner (Section 1.6.4). Statistical thermodynamics uses similar considerations to enable the calculation of thermodynamic values for macroscopic systems from molecular energies. Systems are described in terms of their different states and the probability of finding them in a particular state. We have encountered this concept before, when we introduced the statistical interpretation of entropy (see Section 1.5.5).

The central element of statistical thermodynamics is the partition function, a function that describes the number of states a system can access because of its thermal energy. From the partition function of a system, its internal energy, heat capacity, enthalpy, entropy, Gibbs free energy, and equilibrium constants can be calculated. Partition functions provide a basis to analyze structural transitions in biomolecules (Section 3.3), the assembly of multi-component complexes, or the binding of multiple ligands to enzymes and nucleic acids (Section 3.4).

3.1 PROBABILITIES, STATISTICAL WEIGHTS, ENERGY, AND THE BOLTZMANN DISTRIBUTION

This Section introduces the basic concepts of statistical thermodynamics. We start our considerations from a simple, binary equilibrium between two states, A and B (see Section 2.1). For this system, we have defined the equilibrium constant K_{eq} as

$$K_{eq} = \frac{[B]}{[A]} \qquad \text{eq. 3.1}$$

The equilibrium constant describes the behavior of a large ensemble of molecules. The probability P_A to find molecules of this ensemble in state A is given by

$$P_A = \frac{[A]}{[A]+[B]} \qquad \text{eq. 3.2}$$

while the probability P_B to find molecules in state B is

$$P_B = \frac{[B]}{[A]+[B]} \qquad \text{eq. 3.3}$$

By combining eq. 3.1 with eq. 3.2 or eq. 3.3, we obtain

$$P_A = \frac{1}{1+K_{eq}} \qquad \text{eq. 3.4}$$

and

$$P_B = \frac{K_{eq}}{1+K_{eq}} \qquad \text{eq. 3.5}$$

The sum of these two probabilities is

$$P_A + P_B = 1 \qquad \text{eq. 3.6}$$

which simply means, the molecules must be either in state A or in state B.

We now develop a statistical description of this system. First, we assign a reference state. By convention, the reference state is the state with the lowest free energy. The energies of all states of the system are then defined as the difference to this reference state. Let us assume that in this case, state A is the state with the lower free energy, G_A^0. The relative energy of state A then becomes $\Delta G_A^0 = 0$, the energy of state B is

$$\Delta G_B^0 = G_B^0 - G_A^0 \qquad \text{eq. 3.7}$$

Next, we assign statistical weights to each state of the molecules. The statistical weight expresses the relative concentrations of each state, relative to the concentration of the reference state A. The statistical weights ω for states A and B are then

$$\omega_A = \frac{[A]}{[A]} = 1 \qquad \text{eq. 3.8}$$

and

$$\omega_B = \frac{[B]}{[A]} \qquad \text{eq. 3.9}$$

Note that ω_B is nothing else but the equilibrium constant for the equilibrium between A and B (eq. 3.1). The combination of eq. 3.4 and eq. 3.5 with eq. 3.8 and eq. 3.9 defines the probabilities P_A and P_B in terms of the statistical weights as

$$P_A = \frac{\omega_A}{\omega_A + \omega_B} = \frac{1}{1+\omega_B} \qquad \text{eq. 3.10}$$

and

$$P_B = \frac{\omega_B}{1+\omega_B} \qquad \text{eq. 3.11}$$

Again, the sum of these probabilities is unity, as it should be. We now relate the probabilities to the free energies of these states. From eq. 3.1, eq. 3.7 and eq. 3.9, it follows that

$$\Delta G_B^0 = -RT \ln \omega_B \qquad \text{eq. 3.12}$$

(see also eq. 2.9). By solving eq. 3.12 for ω_B and substituting this expression into eq. 3.10 and eq. 3.11, we now can relate the probability to find a molecule in states A or B to the free energy difference between the states (Figure 3.1):

$$P_A = \frac{1}{1 + e^{-\Delta G_B^0 / RT}} \qquad \text{eq. 3.13}$$

and

$$P_B = \frac{e^{-\Delta G_B^0 / RT}}{1 + e^{-\Delta G_B^0 / RT}} \qquad \text{eq. 3.14}$$

The exponents in eq. 3.14 become zero for $\Delta G_B^0 = 0$, meaning the exponential terms in numerator and denominator have values of unity, hence $P_B = 0.5$. This means that the probability of finding molecules in this system in state B is 0.5 or 50%. Note that $\Delta G_B^0 = 0$ simply means that states A and B have the same free energy. For $\Delta G_B^0 > 0$, we obtain $P_B < 0.5$. This behavior is in accordance with classical thermodynamics, where we learnt that $\Delta G_B^0 > 0$ defines a process that does not occur spontaneously, in this case from A to B. Consistently, for $\Delta G_B^0 < 0$, where the spontaneous reaction is from A to B, we obtain $P_B > 0.5$. However, in this case, by definition, state B with the lower energy would be assigned as the reference state.

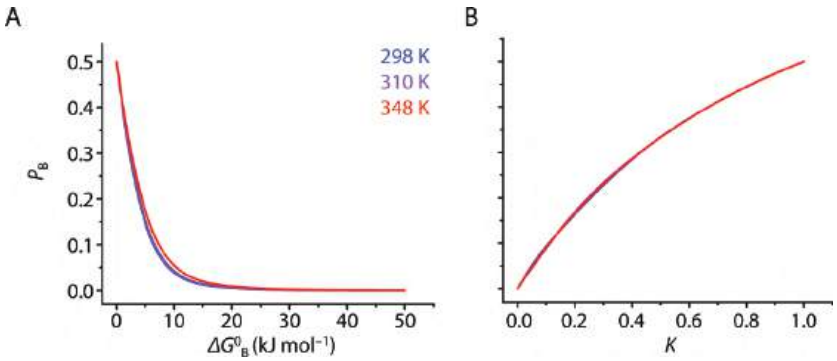


FIGURE 3.1: Energy differences, equilibrium constants, and probabilities. A: The probability P_B to find a molecule in state B as the function of its free energy difference ΔG_B^0 to the reference state A. B: P_B as a function of the equilibrium constant *K*. blue: 298 K, purple: 310 K, red: 348 K.

Note that eq. 3.13 and eq. 3.14 also enable us to derive the ratio of the probabilities of finding molecules in state A or state B as

$$\frac{P_B}{P_A} = \frac{e^{-\Delta G_B^0 / RT}}{1 + e^{-\Delta G_B^0 / RT}} \cdot \frac{1 + e^{-\Delta G_B^0 / RT}}{1} \qquad \text{eq. 3.15}$$

or

$$\frac{P_B}{P_A} = e^{-\Delta G_B^0 / RT} \qquad \text{eq. 3.16}$$

Eq. 3.16 is the *Boltzmann distribution*. It provides a general description of the distribution of an ensemble of molecules over two states divided by a free energy gap ΔG_B^0 (in J mol^{-1}, in accordance with the thermal energy per mol, *RT*, of the ensemble in the denominator of the exponent). We will

put the Boltzmann distribution to use in the treatment of two-state systems in spectroscopy (see Sections 21.1.2, 22.1.1, and 22.2.1). The exponential term $e^{-\Delta G/RT}$, the *Boltzmann factor*, is equivalent to the equilibrium constant (see eq. 2.9), in this case for the equilibrium between state B and the reference state A.

Note that in eq. 3.12–eq. 3.16, we have used ΔG (in J mol^{-1}) to quantify the energies of the system in different states. These expressions thus describe the behavior of a large ensemble of (1 mol of) molecules. We can also formulate these relationships in terms of the energies ε (in J) of individual molecules of the ensemble in states A and B. In this case, we divide the energy difference, $\Delta\varepsilon$, for a single molecule by $k_B T$ instead of RT

$$\frac{P_B}{P_A} = e^{-\Delta\varepsilon/k_B T} \qquad \text{eq. 3.17}$$

where k_B is the Boltzmann constant (see Section 1.5.5, Box 1.8). P_A and P_B then define the probabilities to observe this molecule in states A or B. Note that eq. 3.16 and eq. 3.17 are identical, as both ΔG and R are divided by N_A. We can therefore interpret the probabilities P_A and P_B both on the ensemble level, as the fraction of molecules in the entire system at any given time, and on the molecular level, as the probability to find a particular molecule of the system in state A or B at a given time. In fact, the ratio also tells us how much time any single molecule within the system spends in states A and B. The fact that the time an individual molecule of the system spends in each state over time recapitulates the distribution of a large ensemble of molecules over these states at any given point in time is the principle of *ergodicity*. We will revisit this principle in the context of single-molecule analyses (see Sections 12.2, 25.7.2) and molecular dynamics simulations (Section 20.1.5.3).

Note that in eq. 3.16 and eq. 3.17, the exponent approaches $-\infty$ for $T \to 0$, which means P_B/P_A approaches zero: the molecule(s) almost exclusively populate(s) state A, the *ground state*. For $T \to \infty$, the exponent approaches zero, and P_B/P_A approaches a value of unity: both states are populated equally. This temperature dependence becomes relevant in spectroscopy (Section 21.1.2).

3.2 SUM OF STATES: THE PARTITION FUNCTION

3.2.1 General form of the partition function

The *partition function* or "sum of states" is a central element of statistical thermodynamics that links the energy levels of individual molecules to the thermodynamic behavior of ensembles. It describes the number of states that are accessible to a thermodynamic system at a given temperature T. The partition function is the sum of the statistical weights of all states. For the binary system discussed in Section 3.1, the partition function Z is

$$Z = \omega_A + \omega_B \qquad \text{eq. 3.18}$$

or

$$Z = 1 + e^{-\Delta G_B^0/RT} \qquad \text{eq. 3.19}$$

For a more complex system with S possible states, the partition function is simply the sum of the statistical weights ω_j of all states j of the system. The general form of the partition function then becomes

$$Z = \sum_{j=1}^{S} \omega_j \qquad \text{eq. 3.20}$$

Again, we define the state of the system with the lowest energy as a reference state, and express all other energies ΔG_j^0 relative to this reference state. This leads to the general form of the partition function in terms of energies as

$$Z = \sum_{j=1}^{S} e^{-\Delta G_j^0 / RT} \qquad \text{eq. 3.21}$$

The individual exponential terms in eq. 3.21 are the Boltzmann factors of each state (Section 3.1). Note that the partition function tells us how many states a system can access (thermally) at a given temperature (Figure 3.2). For $T \to 0$, the exponent in the individual Boltzmann factors of the partition function in eq. 3.21 becomes $-\infty$, and all elements except the one for the lowest-energy state approach a value of zero, leaving $Z = 1$: the system can only access one state, the state with the lowest energy (ground state). For $T \to \infty$, the exponents approach zero, and all Boltzmann factors approach a value of one, giving $Z = S$: all possible states are occupied. More generally, the terms with $\Delta G_j^0 \ll RT$ contribute a value of unity, those with $\Delta G_j^0 \gg RT$ approach zero and do not contribute to Z. Hence, the partition function tells us how the energy of the system is distributed, which establishes the link to thermodynamic information (see Sections 1.5.5, 1.6.4). This property of the partition function also provides a link to quantum mechanics and spectroscopy: quantum mechanics describes the possible states of a molecule and their energy levels, spectroscopy (see Section 21.1.2) induces transitions between these states.

Finally, the probability P_i to find a system with S possible states in a particular state i is the ratio of the statistical weight of state i and the sum of the statistical weights of all possible states, *i.e.*

$$P_i = \frac{\omega_i}{Z} = \frac{\omega_i}{\sum_j^S \omega_j} \qquad \text{eq. 3.22}$$

This expression is equivalent to the Boltzmann factor of state i divided by the sum of all Boltzmann factors:

$$P_i = \frac{e^{-\Delta G_i^0 / RT}}{\sum_j^S e^{-\Delta G_j^0 / RT}} \qquad \text{eq. 3.23}$$

Again, the sum of all probabilities P_i has a value of unity. Note that we can re-write eq. 3.23 as

$$P_i = \frac{K_i}{Z} \qquad \text{eq. 3.24}$$

where K_i denotes the equilibrium constant between the state i and the reference state. This is consistent with the definition of the statistical weight of state i (eq. 3.9).

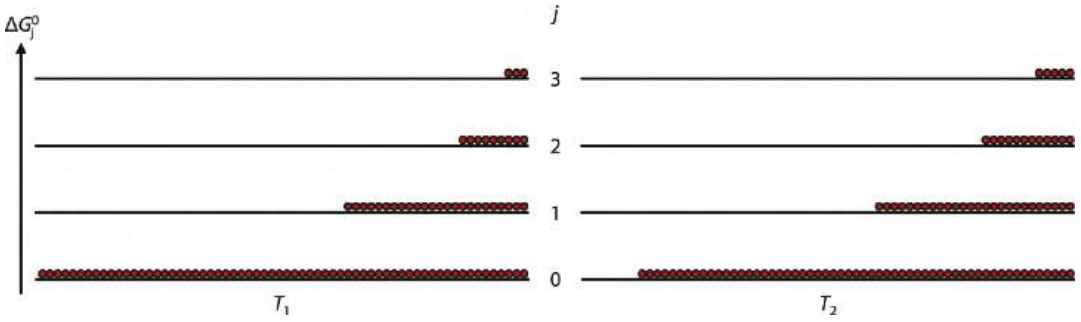


FIGURE 3.2: Boltzmann factors and probabilities. The distribution of 100 particles over 4 states with free energies zero (reference state, $j = 0$), RT_1 ($j = 1$), $2{\cdot}RT_1$ ($j = 2$), and $3{\cdot}RT_1$ ($j = 3$) at $T_1 = 277$ K (left) and $T_2 = 373$ K (right), calculated according to eq. 3.23. With increasing temperatures, the higher-energy states become more populated due to the increased thermal energy.

3.2.2 The partition function for systems with degenerate states

Note that eq. 3.20 to eq. 3.23 hold only for a system with S states of different energies. Sometimes, different states have the same energy. This happens if states can be generated in different ways. For example, a dimeric enzyme with two active sites can exist in the free state, without any ligands bound, in a state with one ligand bound, or in a state with two ligands bound. These states are the macrostates of the system (see Section 1.5.5). While there is only one way to generate an enzyme with two free binding sites or with both binding sites occupied, the state with one binding site occupied can be generated in two different ways, by binding of a ligand to site 1 or to site 2, leading to two microstates of this macrostate (see Sections 1.5.5 and 13.4).

If the two binding sites are identical, the two microstates cannot be distinguished. They have the same energy; we then say this macrostate is degenerate, with a *degeneracy* of $g = 2$. The degeneracy g_j of each macrostate j of the system has to be taken into account in the partition function, which becomes

$$Z = \sum_{j}^{S} g_j \omega_j \qquad \text{eq. 3.25}$$

or

$$Z = \sum_{j}^{S} g_j e^{-\Delta G_j^0 / RT} \qquad \text{eq. 3.26}$$

The only difference to the general form of the partition function for a system with S states of unique energies (eq. 3.21) is the weighting of the Boltzmann factors for each state by its degeneracy. The probability P_i to find a system with degeneracy in a particular state i then becomes

$$P_i = \frac{g_i \omega_i}{\sum_{j}^{S} g_j \omega_j} \qquad \text{eq. 3.27}$$

or, in terms of energies (Figure 3.3):

$$P_i = \frac{g_i e^{-\Delta G_i^0 / RT}}{\sum_{j}^{S} g_j e^{-\Delta G_j^0 / RT}} \qquad \text{eq. 3.28}$$

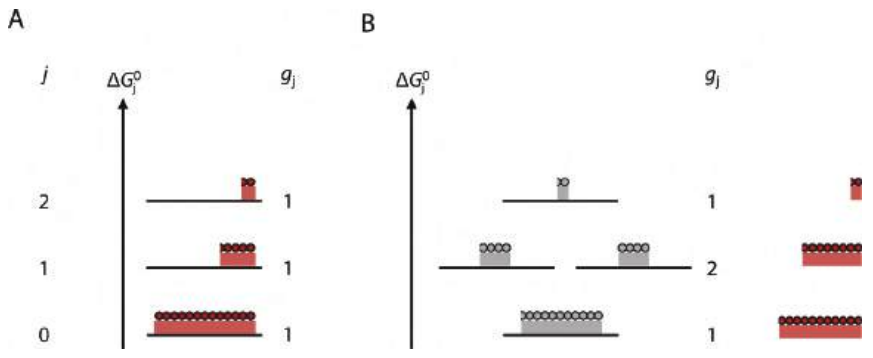


FIGURE 3.3: Effect of degeneracy on probabilities. A: Distribution of molecules over three states. B: Distribution of molecules over a system with three states of different energy, and a degeneracy $g = 2$ for the intermediate state ($j = 1$). The gray bars reflect the probabilities of finding a molecule in the respective state, the red bars are the sum over all states with the same energy. Distributions are calculated after eq. 3.28 with $T = 298$ K, and free energies of the three states are zero ($j = 0$), RT ($j = 1$), and $2{\cdot}RT$ ($j = 2$).

3.2.3 From the molecular partition function to the partition function of ensembles

The partition function sums up the Boltzmann factors of all possible states of a system, weighted by the degeneracy of each state (eq. 3.26). The individual terms that enter into the partition function of a system depend on the molecules that constitute this system, their energy, and, if present, the energy of their interactions with each other.

In the following, we link the properties of the molecules to the properties of the ensemble. We will restrict our considerations to ensembles of non-interacting molecules. To distinguish the molecular partition function from the partition function Z of an ensemble, we denote it with a lower case "z". In analogy to eq. 3.21, the general form of the molecular partition function z for a molecule with S states is

$$z = \sum_{j}^{S} e^{-\varepsilon_j / k_B T} \qquad \text{eq. 3.29}$$

where ε_j denotes the energy of state j relative to the energy of the reference state (ground state) of the molecule. The molecular partition function z contains contributions from translational (trans), rotational (rot), and vibrational energy (vib) of the individual molecule, and from its electronic energy (el). If these contributions are independent of each other, the energy ε_j of each molecule is the sum of the different energy terms:

$$\varepsilon_j = \varepsilon_j^{\text{trans}} + \varepsilon_j^{\text{rot}} + \varepsilon_j^{\text{vib}} + \varepsilon_j^{\text{el}} \qquad \text{eq. 3.30}$$

Substituting this expression into eq. 3.29 gives

$$z = \sum_{j}^{S} e^{-(\varepsilon_j^{\text{trans}} + \varepsilon_j^{\text{rot}} + \varepsilon_j^{\text{vib}} + \varepsilon_j^{\text{el}})/kT} \qquad \text{eq. 3.31}$$

or

$$z = \sum_{j}^{S} e^{-\varepsilon_j^{\text{trans}}/k_B T} \cdot \sum_{j}^{S} e^{-\varepsilon_j^{\text{rot}}/k_B T} \cdot \sum_{j}^{S} e^{-\varepsilon_j^{\text{vib}}/k_B T} \cdot \sum_{j}^{S} e^{-\varepsilon_j^{\text{el}}/k_B T} \qquad \text{eq. 3.32}$$

Thus, the molecular partition function z is the product of the individual partition functions for the translational, rotational, vibrational, and electronic states:

$$z = z^{\text{trans}} \cdot z^{\text{rot}} \cdot z^{\text{vib}} \cdot z^{\text{el}} \qquad \text{eq. 3.33}$$

These individual partition functions depend on the mass of the molecule and its accessible volume (z^{trans}), and on its structure (z^{rot}, z^{vib}, and z^{el}), which defines the energy levels of the different possible rotational, vibrational, and electronic states. If the energy levels of the different states involved are known, either from *ab initio* calculations for simple molecules or from experimental (spectroscopic) data for more complex molecules, the molecular partition function can be calculated explicitly. However, this calculation is beyond the scope of this textbook; the interested reader is referred to more specialized literature.

Knowing the molecular partition function for the individual molecules, we can now return to the partition function for the entire ensemble. For an ensemble of n identical, non-interacting molecules, the energy of the ensemble is simply the sum of the energies of the individual molecules:

$$E_j = \varepsilon_j^1 + \varepsilon_j^2 + \varepsilon_j^3 + \ldots \varepsilon_j^n \qquad \text{eq. 3.34}$$

In eq. 3.34, ε_j^1 refers to the energy of molecule 1 in an ensemble in a state j with the energy E_j. The partition function of the ensemble then is

$$Z = \sum_j^S e^{-\left(\varepsilon_j^1 + \varepsilon_j^2 + \varepsilon_j^3 + \ldots \varepsilon_j^n\right)/k_B T} \qquad \text{eq. 3.35}$$

The sum in eq. 3.35 indicates that we have to sum over all states of the system. We can generate each state of the ensemble by letting each individual molecule go through all possible states. Alternatively, we can sum over all states j for each molecule, and then multiply these sums, *i.e.*

$$Z = \sum_j^S e^{-\frac{\varepsilon_j^1}{k_B T}} \cdot \sum_j^S e^{-\frac{\varepsilon_j^2}{k_B T}} \cdot \sum_j^S e^{-\frac{\varepsilon_j^3}{k_B T}} \cdot \ldots \cdot \sum_j^S e^{-\frac{\varepsilon_j^n}{k_B T}} \qquad \text{eq. 3.36}$$

As the molecules are identical, the superscripts from 1 to n in eq. 3.36 are not needed, and the expression simplifies to

$$Z = \left(\sum_j^S e^{-\frac{\varepsilon_j}{k_B T}} \right)^n \qquad \text{eq. 3.37}$$

Before we simplify further, we have to remember that identical particles have the same chemical identity. However, they can be distinguishable or indistinguishable. In crystals or solids, where the individual particles cannot exchange their places freely, they can be distinguished by their position. In gases, where the particles can freely exchange their places, they cannot be distinguished.

For an ensemble of indistinguishable molecules, we need to take into account that the states of the ensemble can be generated in different ways. Let us consider a tiny ensemble with two identical, indistinguishable molecules that can be in two different states, a low-energy and a high-energy state. The entire system then has three possible (macro-)states: state 1 with both molecules in a low energy state, state 2 with one molecule in a low-energy state, the other in a high-energy state, and state 3 with both molecules in a high-energy state. In eq. 3.37, state 2, with a degeneracy of two, would be counted twice, once for the (micro-)state in which molecule 1 is in the high-energy state, and once for the (micro-)state in which molecule 2 is the one in the high-energy state. However, with indistinguishable molecules, these two microstates cannot be distinguished, only the macrostate should be counted. For n indistinguishable particles, we therefore have to divide the expression in eq. 3.37 by the number of microstates, which is quantified by $n!$ (Section 31.3):

$$Z = \frac{z^n}{n!} \qquad \text{eq. 3.38}$$

For ensembles with n identical, but distinguishable particles, on the other hand, we do need to take into account every single microstate of a particular macrostate, and eq. 3.37 simplifies to

$$Z = z^n \qquad \text{eq. 3.39}$$

With eq. 3.38 and eq. 3.39, we have now established the link between energetic states of a molecule and the states accessible to an ensemble of these molecules. From the partition function, we can calculate the thermodynamic parameters of the ensemble, including the internal energy U, the heat capacity C, the enthalpy H and the entropy S, the Gibbs free energy G, as well as equilibrium constants. For these calculations, we refer the reader to textbooks dedicated to statistical thermodynamics. Instead, we will now illustrate how the concepts of statistical thermodynamics apply to biological systems.

3.3 STATISTICAL THERMODYNAMICS AND STRUCTURAL CHANGES IN BIOMOLECULES

In the following, we will relate the general considerations from Sections 3.1 and 3.2 to biomolecules. Many biomolecules undergo structural transitions that involve two different states. Examples are the formation of an α-helix or β-sheet from a random coil in proteins (see Sections 18.2 and 18.5), the annealing of two single-strands of DNA to a duplex (Section 19.1.1), the transition from B-form to A-form DNA (Section 19.5.1), or phase transitions of lipids (Section 18.4.3). In the following, we will derive the general description of such two-state systems from the perspective of statistical thermodynamics.

3.3.1 General description of structural transitions

We denote the two possible conformations of a biomolecule with n residues as A and B and consider a structural transition from the lower-energy state A, our reference state, to a higher-energy state B. For the biomolecule in state A, all residues are in the conformation *a*. The molecule in state B has all residues in conformation *b*. The two-state transition between states A and B can occur in three fundamentally different ways (Figure 3.4): In the extreme case of an *all-or-none transition*, only state A (all residues in conformation *a*) or state B (all residues in conformation *b*) exist; intermediates between these states are not populated. In a *non-cooperative transition*, each residue can undergo a transition from conformation *a* to conformation *b*, independent of the other residues. In this case, all intermediates with all possible arrangements of residues in conformations *a* or *b* are populated. And finally, in a *zipper* transition, the transition of a residue from conformation *a* to conformation *b* is favored next to a residue that has converted to conformation *b* earlier. As a consequence, zipper-like structural transitions can be divided into an *initiation phase* (first transition from *a* to *b*; nucleation) and an *extension phase* (propagation).

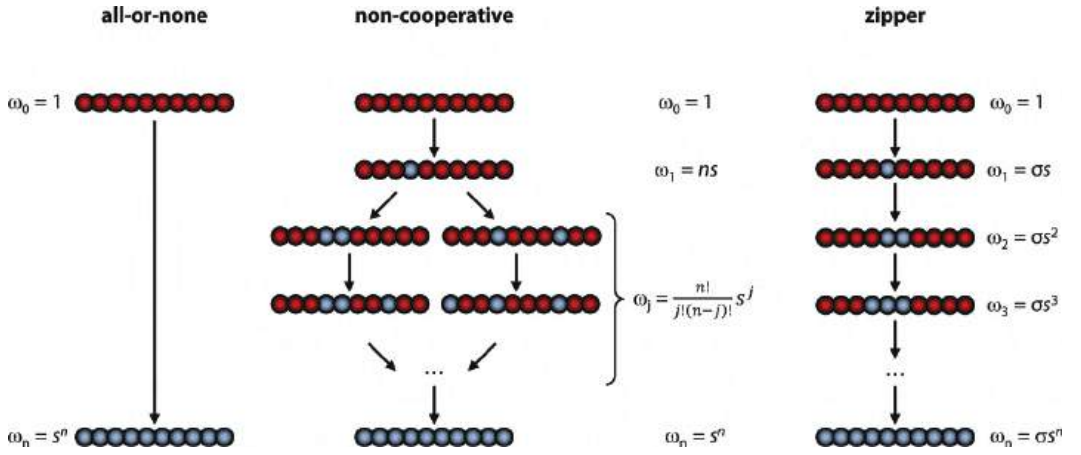


FIGURE 3.4: Schematic of two-state structural transitions. The spheres symbolize individual residues out of a larger biomolecule; the colors indicate two different conformational states. In an all-or-none transition (left), only states with all residues in the same conformational state (red or blue) are possible. In a non-cooperative transition (center), each residue can convert from one to the other state independently of the other residues. In the zipper model (right), the structural transition starts with a nucleation event. In subsequent steps, only residues next to a residue that has converted before undergo a structural transition. The statistical weights ω_j for the entire biomolecule in each state, with j residues converted (from red to blue), are indicated. σ denotes the nucleation parameter, s is the statistical weight of each residue (also termed propagation parameter in the zipper model), n is the number of residues.

To describe these structural transitions using the concepts of statistical thermodynamics, we treat the biomolecule as an ensemble of n residues. Each residue can be in either state *a* or state *b*; the entire molecule can be in macrostates with all-*a* conformation (state A), with one residue in *a* and

n–1 residues in *b*, … , with *n*–1 residues in *a* and one residue in *b*, or with all-*b* conformation (state B). The macrostates of the molecules with residues in different conformations can be generated in different ways.

We start with the all-or-none model. We denote the statistical weight of the only two states possible with ω_0 (no residue in conformation *b*) and ω_n (*n* residues in conformation *b*), and then assign a statistical weight *s* to each residue

$$s = \frac{[b]}{[a]} \qquad \text{eq. 3.40}$$

which describes the equilibrium constant for the transition of this residue from conformation *a* to *b*. The statistical weight ω_n for the entire biomolecule is the product of the individual statistical weights of each residue, and hence

$$\omega_n = \prod_{j=1}^{n} s_j = s^n \qquad \text{eq. 3.41}$$

The partition function for the all-or-none model is thus

$$Z = \omega_0 + \omega_n = 1 + s^n \qquad \text{eq. 3.42}$$

Note that eq. 3.42 is equivalent to eq. 3.19 for a two-state system of states A and B, here with the statistical weights ω_0 and ω_n. The probability to observe the system in state B is

$$P_B = \frac{s^n}{1 + s^n} \qquad \text{eq. 3.43}$$

In the non-cooperative model, the starting point and reference state is the molecule in state A, with all residues in conformation *a*. The statistical weight of the reference state is ω_0. All other states, generated by the stepwise conversion of individual residues from conformation *a* to *b*, have a statistical weight ω_j. Again, we describe the probability for the conversion of a residue from *a* to *b* by a statistical weight *s* (eq. 3.40). The transition can now be regarded as a set of equilibria for each residue, with

$$s = e^{-\Delta G^0_{b-a}/RT} \qquad \text{eq. 3.44}$$

In the absence of cooperativity, the probability for the conversion from *a* to *b* is identical for each and every residue. The statistical weight for the biomolecule with *j* out of *n* residues converted from *a* to *b* is

$$\omega_j = s^j \qquad \text{eq. 3.45}$$

Eq. 3.45 holds for indistinguishable residues. However, biopolymers, *i.e.* proteins or nucleic acids, have a direction (N- to C-terminus, 5' to 3'). Therefore, residues can be distinguished by their position. A polymer of *n* residues with one residue converted from *a* to *b* can exist in *n* different forms, with the residue in conformation *a* in one of the *n* positions. For this state, ω_1 is the product of *n* and *s*. The resulting degeneracy g_j needs to be taken into account for each state, here as part of ω_j:

$$\omega_j = g_j s^j \qquad \text{eq. 3.46}$$

To determine the degeneracy g_j, we count the number of ways in which a state *j* of a macromolecule with *n* residues, out of which *j* residues have converted from *a* to *b*, and *n*–*j* residues are still in

conformation *a*, can be generated. This number is defined by the binomial coefficient (Sections 3.2.2 and 31.3):

$$g_j = \frac{n!}{j!(n-j)!} \qquad \text{eq. 3.47}$$

Substituting eq. 3.47 into eq. 3.46 gives

$$\omega_j = \frac{n!}{j!(n-j)!} s^j \qquad \text{eq. 3.48}$$

The partition function, the sum of all statistical weights, is then

$$Z = \sum_j^n \omega_j = \sum_j^n \frac{n!}{j!(n-j)!} s^j \qquad \text{eq. 3.49}$$

which is equivalent to the polynomial

$$Z = (1+s)^n \qquad \text{eq. 3.50}$$

(Section 31.3). Together with eq. 3.48, we can then express the probability of observing a state *i* as

$$P_i = \frac{\omega_i}{Z} = \frac{\frac{n!}{i!(n-i)!} s^i}{(1+s)^n} \qquad \text{eq. 3.51}$$

Eq. 3.51 tells us that for $s = 1$ (*i.e.* $\Delta G^0_{b-a} = 0$) the state with the highest probability is the state that shows the highest degeneracy, and hence is associated with the highest entropy (see Section 1.5.5).

We can now calculate the fraction f_j of residues in conformation *b* over the entire system. The number of residues in conformation *b* equals *j*, which gives

$$f_j = \frac{j}{n} \qquad \text{eq. 3.52}$$

For the entire system, this gives the probability to observe a residue in conformation *b* as

$$P_b = \frac{\sum_j^n f_j \omega_j}{Z} \qquad \text{eq. 3.53}$$

By expressing the sums in the numerator and the denominator as polynomials (Section 31.3), we obtain

$$P_b = \frac{s \cdot (1+s)^{n-1}}{(1+s)^n} = \frac{s}{1+s} \qquad \text{eq. 3.54}$$

The probability to observe *a* must be $1-P_b$, which is

$$P_a = 1 - P_b = \frac{1}{1+s} \qquad \text{eq. 3.55}$$

By dividing eq. 3.54 by eq. 3.55, we then obtain

$$\frac{P_b}{P_a} = s \qquad \text{eq. 3.56}$$

The ratio of the probabilities to observe states *a* or *b* over the entire biomolecule is thus identical to the equilibrium constant *s* for each residue.

Finally, we come to the zipper model. Again, the structural transition occurs in separate steps. The main characteristic of the zipper model is that the first structural conversion, the initiation or nucleation step, is energetically more costly than the conversion of adjacent residues, the propagation step. In comparison to the non-cooperative model, we need to take into account this energetic cost of the first step, and its resulting lower probability. We do this by including a *nucleation parameter* σ in the statistical weight of state 1:

$$\omega_1 = \sigma s \qquad \text{eq. 3.57}$$

We then need to distinguish between two different possible subsequent steps: if the next step is the conversion of an adjacent residue, then this occurs with the probability *s*:

$$\omega_2 = \sigma s^2 \qquad \text{eq. 3.58}$$

But if the second step is the conversion of an isolated residue elsewhere in the sequence, then the nucleation energy needs to be considered again. In this case,

$$\omega_2 = \sigma^2 s^2 \qquad \text{eq. 3.59}$$

The parameter *s* is also termed the *propagation parameter*.

The *Zimm-Bragg model* of α-helix formation (Section 18.5.5.1) allows non-contiguous regions in the macromolecules, such as …*abbaaabbba*..., generated by separate nucleation events. However, for helix-coil transitions in biopolymers, the value of σ is typically much smaller than unity: $\sigma \ll 1$. In these cases, we can ignore additional nucleation events, and describe the structural transition by a single nucleation event, followed by propagation steps. Each propagation step then introduces an additional factor *s* into the statistical weight, leading to

$$\omega_j = \sigma s^j \qquad \text{eq. 3.60}$$

What is the degeneracy g_j of each state *j*? Again, we need to consider the number of ways this state can be generated (Section 31.3), in this case the number of ways a segment of *j* adjacent residues can be put into a chain of *n* residues, which is

$$g_j = n - j + 1 \qquad \text{eq. 3.61}$$

The reference state, with all residues in conformation *a*, has a degeneracy $g_0 = 1$. We then obtain the partition function *Z* as

$$Z = 1 + \sum_{j=1}^{n} g_j \omega_j = 1 + \sum_{j=1}^{n} (n - j + 1) \sigma s^j \qquad \text{eq. 3.62}$$

or, after factoring out σ:

$$Z = 1 + \sigma \sum_{j=1}^{n} (n - j + 1) s^j \qquad \text{eq. 3.63}$$

The probability to observe a residue in conformation *b* then becomes

$$P_b = \frac{\sum_{j=1}^{n} f_j g_j \omega_j}{Z} = \frac{\sigma \sum_{j=1}^{n} \frac{j}{n}(n-j+1)s^j}{Z} \qquad \text{eq. 3.64}$$

where f_j is again the fraction of residues in conformation b in state j (eq. 3.52), and g_j is the degeneracy. The cooperativity of the zipper transition depends on the values for σ and s. For $\sigma = 1$, the first transition is equally probable as any subsequent transition from a to b: the nucleation step and the propagation steps are assigned equal contributions to the statistical weight, and the transition is non-cooperative. For $\sigma \ll s$, the nucleation step immediately leads to the conversion of all other residues from a to b, and the transition becomes highly cooperative.

3.3.2 Applications of the zipper model to structural transitions of biomolecules

The zipper model applies to many different processes of macromolecules. The physical meaning of the nucleation and propagation parameters depends on the type of macromolecule and the nature of the structural transition. In the following, we will use helix formation in polypeptides, melting and annealing of nucleic acid duplexes, and helical transitions in nucleic acids as illustrative examples.

3.3.2.1 Coil-helix transition in proteins

We can describe the transition of a polypeptide from a random coil to an α-helical structure using the principles of statistical thermodynamics. The canonical α-helix is stabilized by sequential hydrogen bonds between the carbonyl function of residue i and the amide function of residue i+4 (see Section 18.3.1.1). Their formation can be described by the zipper model (Section 3.3.1), with a nucleation process and subsequent propagation steps leading to an extension of the helix. What is the molecular difference between nucleation and propagation steps? During nucleation, five residues need to adopt a local geometry characteristic of α-helices in a way that the first hydrogen bond between residue i and i+4 can be formed (see Section 18.3.1.1). This step is entropically highly disfavored, but enthalpically favored due to the formation of the hydrogen bond. Any subsequent conversion event will directly lead to the formation of an additional hydrogen bond that enthalpically stabilizes the extending helix. Extension steps are also entropically disfavored, but less so, because only a single residue needs to be restrained in its conformation. The higher entropic cost for formation of the first turn with the first hydrogen bond is the molecular reason for the small value of the nucleation parameter σ, which is on the order of 10^{-3} or 10^{-4}, depending on the polypeptide sequence. Values for the propagation parameter s range between 0.6 and 1.2, depending on the type of residue and its helix-forming propensity (see Section 18.2, Table 18.5). Proline has an exceptionally low value of $s = 0.19$. There are two reasons for this: the imido form restricts the conformational freedom of proline, and fixes it in a conformation that is not compatible with α-helix geometry, and prevents the formation of a stabilizing hydrogen bond (see Sections 18.1.1 and 18.3.1.1).

3.3.2.2 Melting and annealing of nucleic acids

Structural changes in nucleic acids, such as transitions between different helix conformations (see Section 19.5.1) or the annealing of single strands and melting of duplex regions (Sections 19.5.1 and 19.6.1), can also be described using the zipper model. To understand the underlying processes during melting and annealing, we need to take into account two main interactions, hydrogen bonds between complementary bases of a base pair (Section 19.1.1) and stacking interactions between adjacent bases (Section 19.4.1). The effect of melting and annealing on these interactions is different, leading to different pathways (Figure 3.5): Melting involves unstacking of adjacent nucleobases and breaking of hydrogen bonds that stabilize the melting base pair. The energetic contribution for the disruption of hydrogen bonds is equal for the first base pair (nucleation) and any subsequent base pair (propagation.) The difference lies with the stacking interactions: to initiate duplex melting, the stacking interactions on both sides of the melting base pair are disrupted. In each subsequent step, the stacking interaction to only one base pair is lost. The nucleation parameter σ accounts for the

loss of these two stacking interactions on melting the first base pair, plus the loss of the stabilizing hydrogen bonds. The propagation parameter *s*, on the other hand, only contains contributions for the disruption of one stacking interaction and the hydrogen bonds disrupted.

Annealing and duplex formation is not merely a reversion of melting. First of all, annealing is an intermolecular reaction, in which two strands associate. Consequently, nucleation depends on the concentration of these strands, whereas propagation does not. Due to the concentration dependence of the nucleation step, low concentrations favor intramolecular interactions in each strand over intermolecular interactions leading to annealing of two strands. Both nucleation and propagation depend on the identity of the base pair, with G-C pairs being more stable than A-T pairs (see Section 19.1.1).

Second, the intermediates formed during annealing are those that form rapidly (kinetic control, see Section 15.4). These intermediates will dissociate and rearrange during the annealing process, until eventually the most stable duplex is formed (thermodynamic control, see Section 15.4). As a consequence, annealing intermediates are more heterogeneous (higher entropy) than intermediates populated during melting (lower entropy). The melting temperature T_m is higher than the annealing temperature T_a, and the cooperativity of melting is higher than the cooperativity for annealing (Figure 3.5).

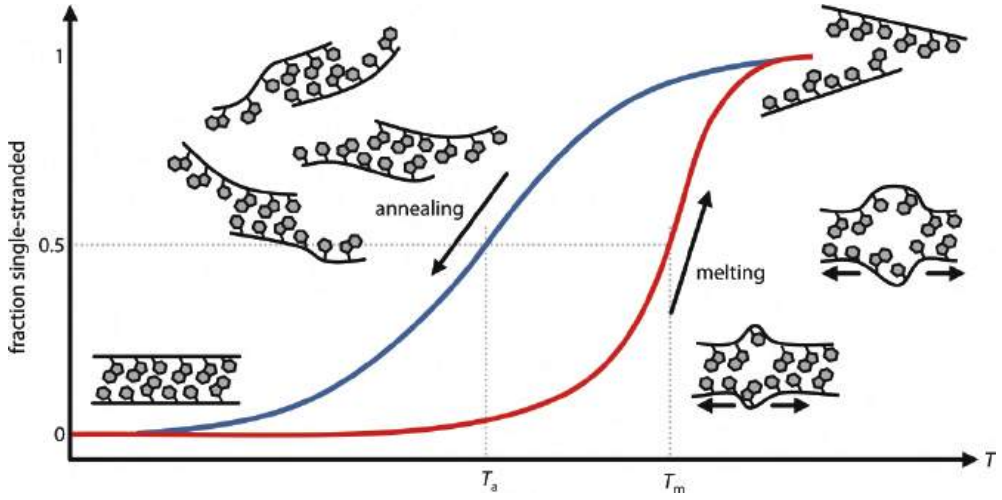


FIGURE 3.5: Different pathways for annealing and melting. DNA melting (red curve) is nucleated in less stable regions of the duplex. The resulting single-stranded bubble then expands in both directions, until the two strands dissociate. Annealing (blue curve) goes through a number of intermediates with partial base-pairing before the most stable, fully base-paired structure is formed. T_m: melting temperature, T_a: annealing temperature.

3.3.2.3 Helix transitions of nucleic acids

The transition of a linear DNA duplex from B-form to A-form geometry (see Section 19.5.1) can be described directly by the zipper model. The propagation parameter *s* depends on the energetic difference between base pairs in both helix geometries. The nucleation parameter σ additionally takes into account the energetic cost of creating a junction between A-form and B-form DNA, which is necessary on both sides of the first converting base pair. For any subsequent conversion, the junction is moved by one base pair, but no additional junction is generated.

The zipper model can also be applied to describe more complex structural transitions, such as transitions between different topological states of DNA, and helical transitions coupled to supercoiling (see Section 19.5.3.2).

3.4 STATISTICAL THERMODYNAMICS AND BINDING EQUILIBRIA

Binding equilibria can be described from a statistical perspective. The statistical treatment comes in handy when applied to larger systems involving the binding of ligands to multiple binding sites in a biomolecule (see Sections 13.4, 13.5). In this case, we consider the ensemble as a set of molecules

with n binding sites; the binding sites can exist in two states: occupied or empty. We can write down the partition function Z for this system as

$$Z = \sum_{j}^{n} g_j \omega_j \qquad \text{eq. 3.65}$$

with the total number of binding sites, n, the number of binding sites occupied, j, and the degeneracy (or ways to generate) a state with j binding sites occupied, g_j. We then use the free state as a reference state and express the partition function in terms of energies of all states involved, which provides a link to the binding energy of the ligand to individual sites and the equilibrium constants of the binding/dissociation equilibria involved.

In the experimental analysis of interactions, ligand binding is typically followed through a spectroscopic probe in titrations of the ligand with the biomolecule or *vice versa* (see Section 21.5.5.1). The spectroscopic signal of the ligand as a function of the concentration of the biomolecule (or *vice versa*) is normalized to values from zero to one. This binding isotherm reflects the fractional saturation of all binding sites present, denoted with θ:

$$\theta = \frac{\bar{\nu}}{n} \qquad \text{eq. 3.66}$$

n is the number of binding sites present, and $\bar{\nu}$ is the degree of binding, *i.e.* the (average) number of binding sites occupied, varying from zero to n (see Sections 13.1, 13.4, and 13.5).

3.4.1 Multiple equivalent, non-interacting binding sites

To quantify the average number of binding sites occupied, $\bar{\nu}$, we need to divide the concentration of the ligand L bound by the overall concentration of the macromolecule P, which gives

$$\bar{\nu} = \frac{[PL] + 2[PL_2] + 3[PL_3] + \ldots + n[PL_n]}{[P] + [PL] + [PL_2] + \ldots + [PL_n]} = \frac{\sum_{i=1}^{n} i[PL_i]}{\sum_{i=0}^{n} i[PL_i]} \qquad \text{eq. 3.67}$$

with values for $\bar{\nu}$ varying from zero to n. The different species PL_n are connected by the corresponding equilibrium constants. We define the equilibrium association constant K_n for the overall equilibrium between the free ligand L and the free macromolecule P on one side and the complex with n binding sites occupied on the other side as

$$K_n = \frac{[PL_n]}{[P][L]^n} \qquad \text{eq. 3.68}$$

Thereby, we can express eq. 3.67 as

$$\bar{\nu} = \frac{\sum_{i=1}^{n} iK_i[P][L]^i}{\sum_{i=0}^{n} K_i[P][L]^i} \qquad \text{eq. 3.69}$$

In eq. 3.69, K_0 is the equilibrium constant between the reference state and the free state, which are the same, hence K_0 equals unity. By canceling the concentration of the macromolecule, $[P]$, we obtain

$$\bar{\nu} = \frac{\sum_{i=1}^{n} iK_i[L]^i}{\sum_{i=0}^{n} K_i[L]^i} \qquad \text{eq. 3.70}$$

Eq. 3.70 is the *Adair equation* that describes any binding model for binding of multiple ligands to a macromolecule. Depending on the boundary conditions of the system studied, it can be further modified to be applied to macromolecules with equivalent or non-equivalent, independent or dependent (cooperative) binding sites.

In a system with n equivalent, independent binding sites (Figure 3.6), the binding energy and hence the equilibrium constants for each individual binding event of a ligand L to a state PL_{i-1} to form the state PL_i are equal. This equilibrium association constant k, not to be confused with a rate constant, is defined as

$$k = \frac{[PL_i]}{[PL_{i-1}][L]} \qquad \text{eq. 3.71}$$

This constant k is termed the *microscopic equilibrium constant*. In the derivation of the Adair equation (eq. 3.70), we have used overall equilibrium association constants K_i (eq. 3.68), which are *macroscopic equilibrium constants*.

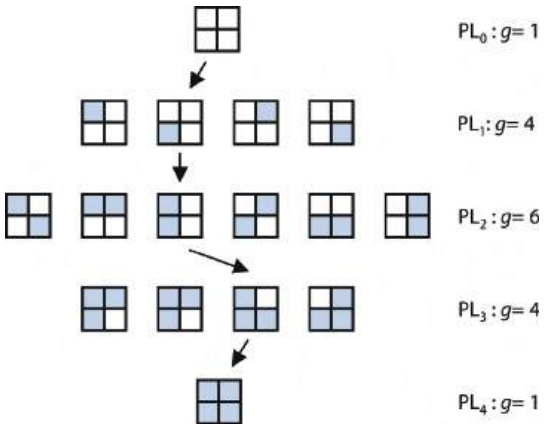


FIGURE 3.6: Binding of ligands to *n* identical, independent binding sites. Ligands can bind to a molecule with four independent binding sites in different ways. The states with no or all binding sites occupied can be generated only in one way, they have a degeneracy of one. The states with one or three binding site(s) occupied have a degeneracy of four, the state with two binding sites has a degeneracy of six. White squares indicate free binding sites, blue squares indicate binding sites occupied with ligand.

The second point we need to consider is the fact that all concentrations $[PL_i]$ refer to macrostates with a certain number i of the n binding sites occupied. These macrostates are degenerate: they can be generated in different ways (Figure 3.6). The concentrations $[PL_i]$ thus equal the sums of concentrations of the individual microstates, in which different sets of i binding sites are occupied. The number of ways each macrostate can be generated, *i.e.* the number of microstates for each macrostate, equals the number of possibilities in which n sites can be split into i occupied sites and $n-i$ unoccupied sites. This value is determined by the binomial coefficients

$$\frac{n!}{(n-i)!i!} \qquad \text{eq. 3.72}$$

(eq. 3.47; see Section 31.3). Tailoring the general form of the Adair equation to this specific binding scenario with these factors then gives

$$\bar{\nu} = \frac{\sum_{i=1}^{n} i \frac{n!}{(n-i)!i!} \left(k[L]\right)^i}{\sum_{i=0}^{n} \frac{n!}{(n-i)!i!} \left(k[L]\right)^i} \qquad \text{eq. 3.73}$$

At first glance, this equation does not look any simpler than the general form of the Adair equation. However, the sums in numerator and denominator can be re-written as binomial expressions. The denominator is

$$\sum_{i=0}^{n} \frac{n!}{(n-i)!i!}\left(k[L]\right)^i = \left(1+k[L]\right)^n \qquad \text{eq. 3.74}$$

For the numerator, we are lacking the term for $i = 0$. We can "compensate" for this by writing the binomial expression with the pre-factor $nk[L]$ as a polynomial of the order n–1 (Section 31.3):

$$\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}\left(k[L]\right)^i = nk[L]\cdot\left(1+k[L]\right)^{n-1} \qquad \text{eq. 3.75}$$

Hence, eq. 3.73 becomes

$$\bar{\nu} = \frac{nk[L]\left(1+k[L]\right)^{n-1}}{\left(1+k[L]\right)^n} \qquad \text{eq. 3.76}$$

or

$$\bar{\nu} = \frac{nk[L]}{1+k[L]} \qquad \text{eq. 3.77}$$

The binding isotherm for n identical, independent binding sites thus is described by a hyperbola that has the same shape as the binding curve for a macromolecule with a single binding site (Figure 3.7). The only difference is the amplitude n. As binding curves are typically measured by using an experimental signal that reports on the relative or fractional saturation θ (eq. 3.66), measured amplitudes are relative, and do not contain information on the number of binding sites present (Figure 3.7; see Section 13.4; eq. 13.47).

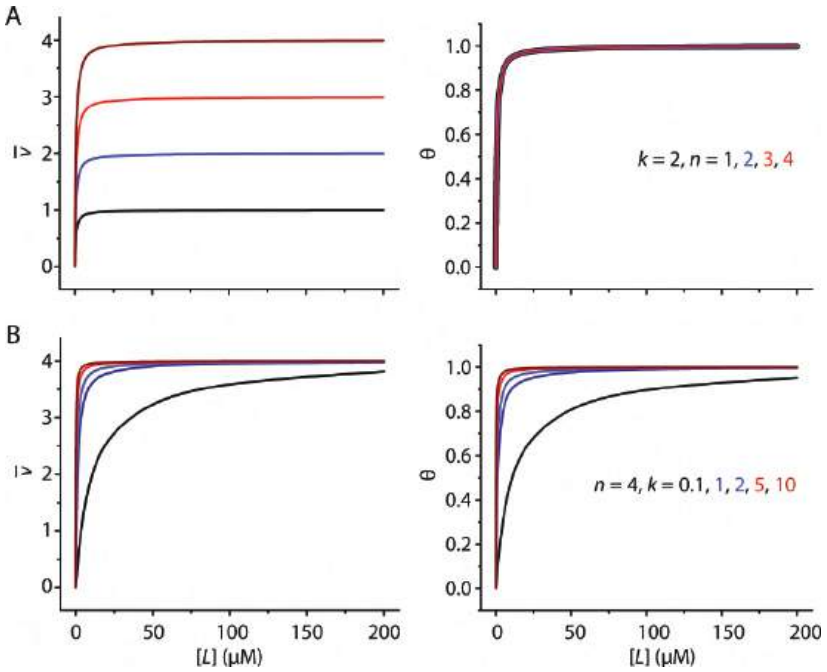


FIGURE 3.7: Binding isotherms for *n* identical, independent binding sites. A: Fraction of binding sites occupied (left) and fractional saturation (right) as a function of ligand concentration for $n = 1$ (black), $n = 2$ (blue), $n = 3$ (red), and $n = 4$ (dark red); the association equilibrium constant is $k = 2$. B: Fraction of binding sites occupied (left) and fractional saturation (right) as a function of ligand concentration for $n = 4$ and $k = 0.1$ (black), $k = 1$ (dark blue), $k = 2$ (blue), $k = 5$ (red), and $k = 10$ (dark red).

3.4.2 Multiple classes of non-equivalent, non-interacting binding sites

Many biomolecules contain a set of non-equivalent binding sites, with different dissociation constants. With the considerations made so far, we can describe this scenario by considering S different classes of binding sites, with a number of n_s binding sites per class. We can then express the fraction $\bar{\nu}$ of binding sites occupied as

$$\bar{\nu} = \sum_{s=1}^{S} \bar{\nu}_s \qquad \text{eq. 3.78}$$

With an equilibrium association constant k_s for binding of ligands to binding sites of class s we obtain

$$\bar{\nu}_s = \frac{n_s k_s [L]}{1 + k_s [L]} \qquad \text{eq. 3.79}$$

The average number of binding sites occupied, $\bar{\nu}$, can be expressed as

$$\bar{\nu} = \sum_{s=1}^{S} n_s \frac{k_s [L]}{1 + k_s [L]} \qquad \text{eq. 3.80}$$

Thus, for a protein with S classes of binding sites, the binding isotherm is a sum of S hyperbolae, weighted by the number of binding sites n_s in each class (Figure 3.8). A common example for two classes of binding sites is the binding of metal ions to nucleic acids. Metal ions bind to numerous sites with low affinity due to electrostatic interactions with the negatively charged (deoxy-)ribose-phosphate backbone (see Section 19.1). In some cases, specific binding sites with direct coordination of metal ions enable high-affinity binding.

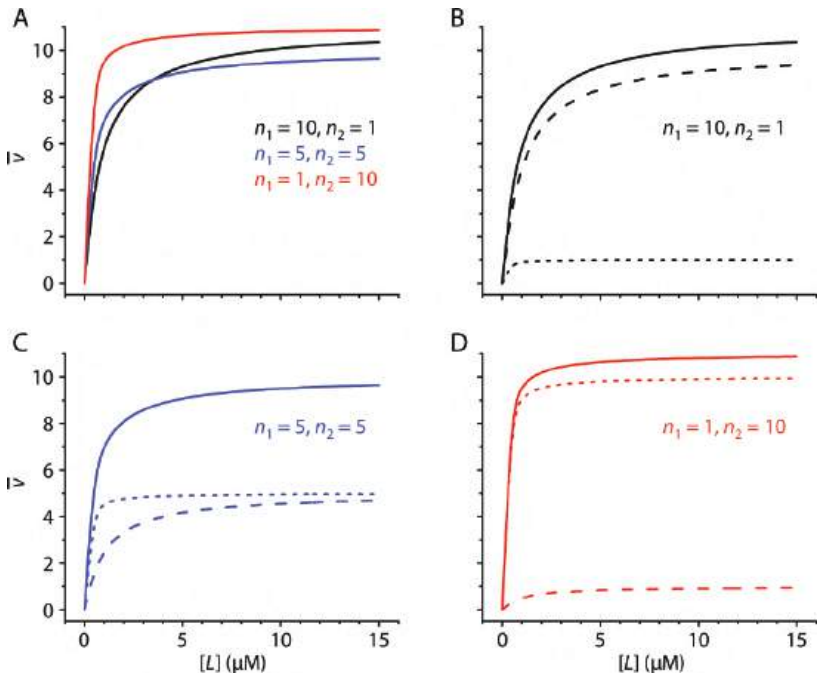


FIGURE 3.8: Binding isotherms for a molecule with two classes of different binding sites. A: Binding isotherms for $k_1 = 1$ (low-affinity binding sites), $k_2 = 10$ (high-affinity binding sites), $n_1 = 10$, $n_2 = 1$ (black), $n_1 = 5$, $n_2 = 5$ (blue), and $n_1 = 1$, $n_2 = 10$ (red). B-D: Binding isotherms from panel A and contributions from the low- and high-affinity binding sites (dashed and dotted lines).

3.4.3 Cooperative binding sites

We can also extend our considerations to the phenomenon of cooperative binding, where the equilibrium association constants of individual binding sites depend on the occupancy of other sites. We have encountered a first example of such cooperativity (thermodynamic linkage) in the context of thermodynamic cycles, where we discussed independent and inter-dependent binding of two substrates A and B to an enzyme E (Section 2.4).

Cooperativity can be described by two fundamentally different models, the Monod-Wyman-Changeux (MWC) model, and the Koshland-Nemethy-Filmer (KNF) model. Both models assume that the binding sites can exist in low-affinity and high-affinity states. In the MWC model the macromolecule exists in only two different states, either with all binding sites in a low-affinity state (T-state, tense), or with all binding sites in a high-affinity state (R-state, relaxed). The equilibrium between these states favors T in the absence of ligand; ligand binding then shifts the equilibrium towards the R-state. It is the concerted transition of all binding states from T to R on binding of just a single ligand that leads to cooperativity. The KNF model, on the other hand, assumes that each binding site can switch between low-affinity and high-affinity states independently of the others, allowing the population of an entire zoo of different species.

3.4.3.1 Monod-Wyman-Changeux model

According to the MWC model, both the T- and the R-states can bind n ligands, and all complexes from R to RL_n and T to TL_n exist (Figure 3.9).

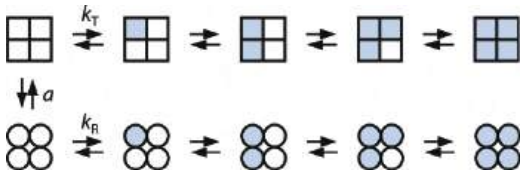


FIGURE 3.9: Monod-Wyman-Changeux model for cooperative ligand binding. Cooperative binding following the MWC model, illustrated for a biomolecule with four independent binding sites. The binding sites exist in two states, a low-affinity state (T, square) and a high-affinity state (R, circle), with an equilibrium constant a in the unbound state. Binding of the ligand to these states is governed by the equilibrium association constants k_T and k_R, with $k_T \ll k_R$. White squares indicate free binding sites, blue squares indicate binding sites occupied with ligand.

In the absence of the ligand, the majority of molecules are in the T-state. The equilibrium constant a is

$$a = \frac{[T]}{[R]} \qquad \text{eq. 3.81}$$

Stronger binding of the ligand to the R-state shifts the equilibrium towards R according to the principle of Le Chatelier (see Section 2.1.2). This conversion of low-affinity sites to unoccupied high-affinity binding sites is what causes the cooperativity. We can now write the average number of binding sites occupied, $\bar{\nu}$, as

$$\bar{\nu} = \frac{\sum_{i=1}^{n} i[RA_i] + \sum_{i=1}^{n} i[TA_i]}{\sum_{i=0}^{n} [RA_i] + \sum_{i=0}^{n} [TA_i]} \qquad \text{eq. 3.82}$$

which is an analogous expression to eq. 3.67. We then define the microscopic equilibrium association constants k_T and k_R in analogy to eq. 3.71, with $c = k_T/k_R$. We also take into account the number of ways in which each state between RL_1 and RL_{n-1} or TL_1 and TL_{n-1} can be generated (eq. 3.72). In analogy to the derivation of eq. 3.73, we replace the concentrations of the ligand complexes of the R- and T-states by the definition of the equilibrium association constants k_R and k_T, and add binomial coefficients to take into account the degeneracy of each complex. Thereby, we obtain

$$\bar{\nu} = \frac{[R]\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}k_R^i[L]^i + [T]\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}k_T^i[L]^i}{[R]\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}k_R^i[L]^i + [T]\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}k_T^i[L]^i} \qquad \text{eq. 3.83}$$

We now substitute $a[R]$ for $[T]$ and ck_R for k_T, which gives

$$\bar{\nu} = \frac{[R]\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}k_R^i[L]^i + a[R]\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}c^i k_R^i[L]^i}{[R]\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}k_R^i[L]^i + a[R]\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}c^i k_R^i[L]^i} \qquad \text{eq. 3.84}$$

or

$$\bar{\nu} = \frac{\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}k_R^i[L]^i + a\sum_{i=1}^{n} i\frac{n!}{(n-i)!i!}c^i k_R^i[L]^i}{\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}k_R^i[L]^i + a\sum_{i=0}^{n}\frac{n!}{(n-i)!i!}c^i k_R^i[L]^i} \qquad \text{eq. 3.85}$$

In analogy to eq. 3.76, we then re-write the individual sums in the numerator and in the denominator as binomials, which leads to

$$\bar{\nu} = nk_R[L]\frac{(1+k_R[L])^{n-1} + ac(1+ck_R[L])^{n-1}}{(1+k_R[L])^n + a(1+ck_R[L])^n} \qquad \text{eq. 3.86}$$

The effect of different values for a, c, and k_R on the average number of binding sites occupied is illustrated in Figure 3.10.

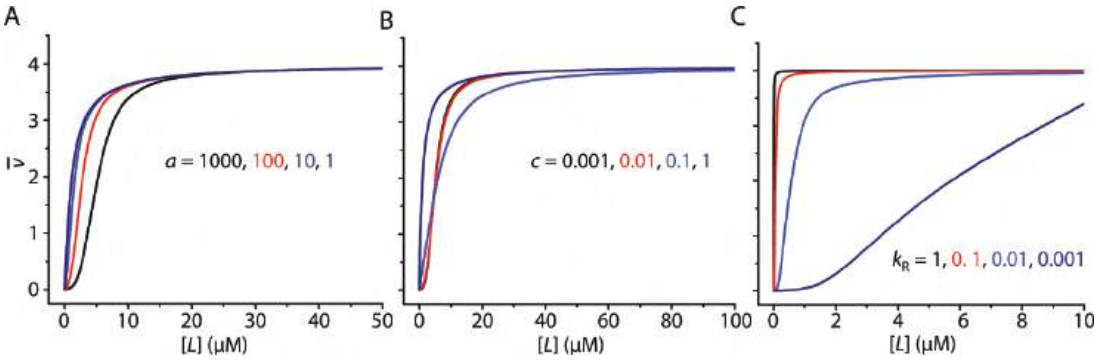


FIGURE 3.10: Cooperative binding and the Monod-Wyman-Changeux model. Effect of the parameters a, c, and k_R. A: $k_R = 1$, $c = 0.001$, $a = 1000$, 100, 10, and 1 (black, red, blue, and dark blue). B: $k_R = 1$, $a = 1000$, $c = 0.001$, 0.01, 0.1, and 1 (black, red, blue, and dark blue). C: $a = 1000$, $c = 0.001$, $k_R = 1$, 0.1, 0.01, and 0.001 (black, red, blue, and dark blue).

3.4.3.2 Koshland-Nemethy-Filmer model

The KNF model does not assume that all binding sites convert in a concerted manner from a low-affinity state to a high-affinity state. Instead, only the binding site engaged with a ligand changes its state, giving rise to mixed species (Figure 3.11). In this model, cooperativity would result from a different interaction between protein subunits that carry the individual binding sites. In principle, we can now follow the same procedure as for the MWC model to derive an equation that relates the average number of binding sites occupied to the ligand concentration. We refrain from this exercise here because this derivation involves a large number of parameters, but return to this model in

Chapter 13 in the context of steady-state enzyme kinetics, where we use an inductive approach to derive the velocity equation for one of the possible scenarios (see Section 13.5).

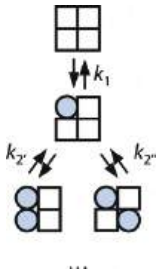


FIGURE 3.11: Cooperative binding and the Koshland-Nemethy-Filmer model. Cooperative binding following the KNF model for a biomolecule with four independent binding sites. The binding sites exist in two states, a low-affinity state (T, square) and a high-affinity state (R, circle). In the absence of ligands, all binding sites are in the T state. While the MWC model assumes that all binding sites convert in a concerted manner from T to R on binding of the first ligand, in the KNF model only the site to which the ligand binds (blue) changes its state. Interactions between two adjacent occupied binding sites (bottom left) and two states across (bottom right) are different.

Statistical thermodynamics can also be applied to more complex interactions, such as intercalation of ligands into DNA (see Section 19.5.4.2). We will return to the statistical perspective on binding equilibria in Chapter 13 in the context of enzyme kinetics, where we derive enzymatic parameters for enzymes with multiple active sites (Sections 13.4 and 13.5).

In this chapter, we have learned how statistical thermodynamics defines the states accessible to a system in terms of the states and associated energies of its constituting molecules. We have seen how these principles can be applied to describe quantitatively biological phenomena, such as structural transitions of biomolecules and binding equilibria.

QUESTIONS

3.1 The entropy of a system is defined as $S = k_B \cdot \ln W$, with W denoting the number of states in which the system can occur. Calculate the entropy S caused by the different states of an enzyme with 1, 2, 4, or 8 binding sites.

3.2 A ligand binds to a protein, the free energy of binding is $\Delta G^{0'} = -22$ kJ mol^{-1}, the enthalpy is $\Delta H^{0'} = 0$ kJ mol^{-1}. Calculate $\Delta S^{0'}$ ($T = 298$ K). What is the molecular interpretation of this behavior?

3.3 A protein exists in a folded and an unfolded state, the free energy of unfolding is 17 kJ mol^{-1} at 25°C. Calculate the probabilities to find the protein in the folded and unfolded state, and use these probabilities to calculate the equilibrium constant of folding.

3.4 A chromophore with a single electronic ground state and a single excited state absorbs light at 280 nm (800 nm). Calculate the partition function at room temperature (298 K) and at 4 K. What do the different values mean? At which temperature do we have 10% of the molecules in the excited state?

3.5 Use the Boltzmann distribution to derive the formula for the decrease in pressure with height.

3.6 Three (distinguishable) systems can be in four states with different energies *0*, *E*, 2*E*, and 3*E*. The total energy is 3*E*. Write down all possible macro- and microstates. Which macrostate is the most probable one and why?

3.7 A peptide in the gas phase has a molecular partition function z. What is the partition function Z of an ensemble of 1 mol of peptide? What is the partition function of a crystal of this peptide with a rectangular unit cell of 1.8 nm x 0.5 nm x 0.7 nm containing two peptides, and a size of 0.27 mm x 0.15 mm x 0.308 mm?

3.8 A peptide consists of four residues that can be either in conformation *a* or *b*. Write down all possible configurations with 0, 1, 2, 3, or 4 residues in *b*, including the statistical weight and degeneracy of each state, and calculate the partition function Z, the probability to find the peptide in state B (*bbbb*), and the average probability to find a residue in conformation *b*, assuming that the transitions from *a* to *b* are non-cooperative, follow the all-or-nothing model, or are described by the zipper model.

3.9 Derive the overall association constants K_1, K_2, and K_3 and dissociation constants K_{d1}, K_{d2}, and K_{d3} for the binding of ligands to three equivalent, independent sites of an enzyme with a microscopic equilibrium association constant k.

3.10 What is the form of the partition function for a macromolecule/ligand system with three identical, independent binding sites?

3.11 Derive the binding isotherm for n identical, independent binding sites (eq. 3.77) in terms of the microscopic dissociation constant k.

3.12 In the original paper by Monod, Wyman, and Changeux, the fractional saturation θ is given as

$$\theta = \frac{Lc\alpha(1+c\alpha)^{n-1} + \alpha(1+\alpha)^{n-1}}{L(1+c\alpha)^{n} + (1+\alpha)^{n}}$$

In this equation, L is the equilibrium constant between T and R ($L = [T_0]/[R_0]$), $c = K_R/K_T$, with K_R and K_T denoting the equilibrium dissociation constants for complexes of R and T, and $\alpha = [L]/K_R$. Show that this equation is equivalent to the expression derived from statistical considerations (eq. 3.86).

3.13 Consider a parallel reaction in which a reactant A is in equilibrium with the reactants B and C. What is the partition function for this system, and what are the probabilities to find the molecule in states A, B, or C? What is the ratio of B and C in equilibrium?

3.14 An enzyme binds two ligand molecules, one with high affinity, the second with low affinity. Calculate the fractional saturation of the low-affinity binding site when the high-affinity binding site is 50% (90%) saturated for K_{d1} and $K_{d2} = 2 \cdot K_{d1}$, for K_{d1} and $K_{d2} = 10 \cdot K_{d1}$, and for K_{d1} and $K_{d2} = 100 \cdot K_{d1}$. What is the overall fraction of binding sites occupied?

3.15 An enzyme exists in an autoinhibited (AI) and an uninhibited, active state (E). The equilibrium constant of autoinhibition, K_{ai}, is

$$K_{ai} = \frac{[AI]}{[E]}$$

Both forms of the enzyme interact with its ligand L. Calculate the apparent dissociation constant of the enzyme-target complexes from statistical thermodynamics considerations. In Chapter 2, the same task is solved using classical thermodynamics.

3.16 A system can exist in two states, separated by the energy difference ΔE. Calculate the temperature at which the higher-energy state is populated to 10% (a) without degeneracy, (b) for a degeneracy of $g = 2$ for the higher-energy state.

REFERENCES

Adair, G. S. (1925) The hemoglobin system. VI. The oxygen dissociation curve of hemoglobin. *J Biol Chem* **63**: 529–545.
· original paper deriving the Adair equation

Altmann, K. H., Wójcik, J., Vásquez, M. and Scheraga, H. A. (1990) Helix-coil stability constants for the naturally occurring amino acids in water. XXIII. Proline parameters from random poly (hydroxybutylglutamine-co-L-proline). *Biopolymers* **30**(1–2): 107–120.
· table for helix propagation parameters for individual amino acids

Atkins, P. and de Paula, J. (2006) *Physical Chemistry for the Life Sciences*. New York, Freeman.
· describes the calculation of thermodynamic parameters from partition functions

Atkins, P. W. (1990) *Physikalische Chemie*. Weinheim, VCH.
· systematic, in-depth treatment of statistical thermodynamics and the calculation of thermodynamic parameters from partition functions

Chay, T. R. and Ho, C. (1973) Statistical mechanics applied to cooperative ligand binding to proteins. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Proc Natl Acad Sci U S A* **70**(12): 3914–3918.
· statistical thermodynamics treatment of the Koshland-Nemethy-Filmer model

Koshland, D. E. Jr., Némethy, G. and Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5**(1): 365–385.
· original reference describing the Koshland-Nemethy-Filmer model of cooperativity

Marzen, S., Garcia, H. G. and Phillips, R. (2013) Statistical mechanics of Monod-Wyman-Changeux (MWC) models. *J Mol Biol* **425**(9): 1433–1460.
· description of Monod-Wyman-Changeux (MWC) models using statistical mechanics

Monod, J., Wyman, J. and Changeux, J. P. (1965) On the nature of allosteric transitions: A plausible model. *J Mol Biol* **12**: 88–118.
· original reference describing the Monod-Wyman-Changeux model of cooperativity

van Holde, K. E., Johnson, W. C. and Ho, P. S. (1998) *Principles of Physical Biochemistry*. Upper Saddle River, Prentice Hall.
· application of the zipper model to complex transitions of biomolecules

Zimm, B. H. and Bragg, J. K. (1958) Theory of the phase transition between helix and random coil in polypeptide chains. *J Chem Phys* **31**: 526–535.
· original reference describing the Zimm-Bragg model of helix-coil transitions



4 Thermodynamics of transport processes

4.1 DIFFUSION

We encountered the spontaneous dissipation of matter in Section 1.5.1. A tendency of particles to achieve a uniform spatial distribution is equivalent to the tendency of molecules in solution to achieve uniform concentrations that are independent of position throughout the solution. Local differences in concentrations are usually equilibrated by diffusion. *Diffusion* is a spontaneous process in three dimensions. If a concentration gradient is present in one dimension, we can define the *flux J* as the number of particles that pass a certain cross-sectional area per time. The magnitude of the flux depends on the concentration gradient. For one-dimensional diffusion in the x-direction, the flux J is described by *Fick's first law*:

$$J = -D\frac{\mathrm{d}c}{\mathrm{d}x} \qquad \text{eq. 4.1}$$

D is the diffusion coefficient (in the x-direction, with the unit $m^2\ s^{-1}$), and dc/dx is the concentration gradient in the x-direction (in $M\ m^{-1}$). For concentrations in particles per volume, the flux is expressed in particles per area and time. For molar concentrations, the flux is expressed in units of $mol\ m^{-2}\ s^{-1}$. The derivation of Fick's first law is illustrated in Box 4.1. For three-dimensional diffusion, the components of the flux in each dimension x, y, and z are described by eq. 4.1. The flux in three dimensions depends on the three-dimensional concentration gradient:

$$\vec{J} = -D\nabla c \qquad \text{eq. 4.2}$$

The inverted triangle in eq. 4.2 is the gradient (Nabla) operator that symbolizes the concentration radients in x-, y-, and z-directions. The negative signs in eq. 4.1 and eq. 4.2 express that the flux is positive for a negative concentration gradient (from high to low concentrations), and negative for a positive concentration gradient (from low to high concentration). Fick's first law thus tells us that macroscopic diffusion in the presence of concentration gradients has a preferred direction:

the molecules diffuse down the steepest gradient. This is in accordance with the concentration dependence of the chemical potential (Section 1.7.5): the higher chemical potential in regions with high concentrations drives the diffusion of molecules toward regions with lower concentration. The flux stops when the concentration is equilibrated over the entire volume, and the driving force for further directed movement is zero. Even when the macroscopic movement stops in equilibrium, the individual molecules do not come to a halt. Due to their thermal energy, molecules can also diffuse in the absence of concentration gradients. This microscopic movement is random, without a preferred direction. Such a random movement is called *Brownian diffusion* or *Brownian motion*. Under certain circumstances, the Brownian motion can be biased to generate directed movement (Section 4.4).

Fick's second law relates the change in concentration with time, dc/dt, to the local concentration gradient dc/dx and has the form

$$\frac{\mathrm{d}c}{\mathrm{d}t} = D\frac{\mathrm{d}^2c}{\mathrm{d}x^2} \qquad \text{eq. 4.3}$$

The derivation of Fick's second law is given in Box 4.1. We will revisit Fick's second law in the context of analytical ultracentrifugation (Section 29.2, eq. 29.26).

BOX 4.1: DERIVATION OF FICK'S LAWS OF DIFFUSION.

We consider a tube with a concentration gradient from one end to the other. The concentration of particles at position x along the tube is $c(x)$ (Figure 4.1). The flux through the cross-section A of the tube at position x can be calculated by accounting for molecules passing through the window from left to right and from right to left in a time window Δt. The average concentration of particles (in particles per volume) in a volume element of length $2b$ around $x-b$ and $x+b$ is $c(x-b)$ and $c(x+b)$. The number of particles n_{left} in this volume element around $x-b$ is

$$n_{\text{left}} = c(x-b)\cdot 2b\cdot A \qquad \text{eq. 4.4}$$

and the number of particles n_{right} in the volume element around $x+b$ is

$$n_{\text{right}} = c(x+b)\cdot 2b\cdot A \qquad \text{eq. 4.5}$$

The number of particles passing through the window at position x in both directions in the time window Δt is proportional to the number of particles in the volume elements left and right and to the time window Δt, giving the flux in number of particles as

$$n_{\text{left}\to\text{right}} = \text{const.}\cdot c(x-b)\cdot 2b\cdot A\cdot \Delta t \qquad \text{eq. 4.6}$$

and

$$n_{\text{right}\to\text{left}} = \text{const.}\cdot c(x+b)\cdot 2b\cdot A\cdot \Delta t \qquad \text{eq. 4.7}$$

The net flux J through the cross-section at position x is proportional to the difference of these numbers, divided by the area A and the time window Δt:

$$J = \text{const.}\cdot\frac{c(x-b)\cdot 2b\cdot A\cdot\Delta t - c(x+b)\cdot 2b\cdot A\cdot\Delta t}{A\cdot\Delta t} \qquad \text{eq. 4.8}$$

(*Continued*)

We can express the concentrations at $x-b$ and $x+b$ as

$$c(x-b) = c(x) - b \cdot \frac{dc}{dx} \qquad \text{eq. 4.9}$$

and

$$c(x+b) = c(x) + b \cdot \frac{dc}{dx} \qquad \text{eq. 4.10}$$

Substituting theses expressions into eq. 4.8 gives the flux J as

$$J = -\text{const.} \cdot \Delta 4b^2 \frac{dc}{dx} \qquad \text{eq. 4.11}$$

Eq. 4.11 corresponds to Fick's first law (eq. 4.1) for $4b^2 \cdot \text{const.} = D$.

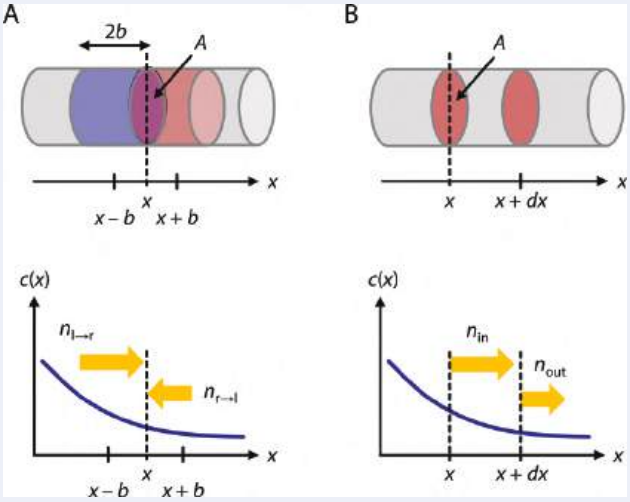


FIGURE 4.1: Derivation of Fick's laws of diffusion. A: To derive Fick's first law, the flux of particles from left to right and from right to left through a window of area A (purple) at position x in a tube with a concentration gradient dc/dx (bottom) is considered. The flux from left to right is proportional to the number of particles in the blue volume element around $x-b$, the flux from right to left is proportional to the number of particles in the red volume element around $x+b$.
B: The derivation of Fick's second law considers the flux of particles into and out of a certain volume element between x and $x+dx$. A net flux occurs only if the concentration gradient is different at x and $x+dx$.

Fick's second law can be derived by similar considerations, starting from a tube with a concentration gradient dc/dx along the x-axis (Figure 4.1). The concentrations at positions x and $x+dx$ along the tubes are $c(x)$ and $c(x+dx)$, with $c(x) > c(x+dx)$. If the concentration gradient at x and $x+dx$ is identical, the flux of particles into the volume element from the left will be identical to the flux of particles from the volume element to the right, and the total flux through this element is zero. If the concentration gradients at x and $x+dx$ are different, the flux in and out of the element differs, and a net flux is observed. For a flux $J(x)$ at position x, we can calculate the number of particles that pass a window of area A at position x during an infinitesimal time interval dt as

(*Continued*)

$$n_{\text{in}} = J(x)A\,\mathrm{d}t \qquad \text{eq. 4.12}$$

and the number of particles that pass this window of area A at position $x+\mathrm{d}x$ is

$$n_{\text{out}} = J(x+\mathrm{d}x)A\,\mathrm{d}t \qquad \text{eq. 4.13}$$

The net flux n_{net} of particles is the difference between these two numbers, giving

$$n_{\text{net}} = J(x)A\,\mathrm{d}t - J(x+\mathrm{d}x)A\,\mathrm{d}t = \big(J(x) - J(x+\mathrm{d}x)\big)A\,\mathrm{d}t \qquad \text{eq. 4.14}$$

The flux $J(x+\mathrm{d}x)$ can be expressed in terms of the flux $J(x)$ and the gradient of the flux, $\mathrm{d}J/\mathrm{d}x$:

$$J(x+\mathrm{d}x) = J(x) + \frac{\mathrm{d}J}{\mathrm{d}x}\cdot\mathrm{d}x \qquad \text{eq. 4.15}$$

Combination of eq. 4.14 and eq. 4.15 gives

$$n_{\text{net}} = -\frac{\mathrm{d}J}{\mathrm{d}x}\cdot\mathrm{d}x\cdot A\cdot\mathrm{d}t \qquad \text{eq. 4.16}$$

We now have to divide by the volume of the element between x and $x+\mathrm{d}x$, $\mathrm{d}V = A\mathrm{d}x$, to convert the change in particle numbers into a change in concentration. The net rate of this change in concentration, $\mathrm{d}c/\mathrm{d}t$, is obtained by dividing by $\mathrm{d}t$. Together, this gives

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -\frac{\mathrm{d}J}{\mathrm{d}x} \qquad \text{eq. 4.17}$$

By expression of the flux J by Fick's first law (eq. 4.1), we obtain

$$\frac{\mathrm{d}c}{\mathrm{d}t} = -\frac{\mathrm{d}\left(-D\dfrac{\mathrm{d}c}{\mathrm{d}x}\right)}{\mathrm{d}x} = D\frac{\mathrm{d}^2c}{\mathrm{d}x^2} \qquad \text{eq. 4.18}$$

which is Fick's second law (eq. 4.3).

Membrane proteins can diffuse in biological membranes because of the membrane fluidity. This diffusion is restricted to the plane of the membrane, and is therefore only two-dimensional (lateral diffusion). The random path that the protein covers during lateral diffusion is called a *trajectory* (Figure 4.2). The *mean square displacement* (MSD) $\langle r^2\rangle$, of the protein can be calculated as

$$\big\langle r^2\big\rangle(\tau) = \Big\langle \big(r(t+\tau) - r(t)\big)^2 \Big\rangle \qquad \text{eq. 4.19}$$

where $r(t)$ is the position of the protein at time t, $r(t+\tau)$ its position at a later time point $t+\tau$, and $r(t+\tau)-r(t)$ is the distance traveled in the time interval τ. The angle brackets denote the time average (from the analysis of a single trajectory) or the ensemble average (from the analysis of a series of trajectories). When analyzing a single trajectory with N data points at intervals Δt, the MSD can be calculated by summing the average distance traveled between all pairs of data points separated by Δt, $2\Delta t$, …, $(N-1)\Delta t$, according to

$$\left\langle r^2 \right\rangle = \frac{1}{N-k}\sum_{i=0}^{N-k}\left(r\left(i\Delta t\right) - r\left(\left(i+k\right)\Delta t\right)\right)^2 \qquad \text{eq. 4.20}$$

where k is an integer number ranging from 1 to (N–1) (Figure 4.2). When an ensemble of n trajectories is analyzed, the MSD is determined by averaging the distance traveled from time t = 0 to the time Δt, $2\Delta t$, …, (N–1)Δt, here averaged over all n trajectories measured (Figure 4.2).

$$\left\langle r^2 \right\rangle = \frac{1}{n}\sum_{i=1}^{n}\left(r_i\left(k\Delta t\right) - r_i\left(0\right)\right)^2 \qquad \text{eq. 4.21}$$

r_i (t) denotes the position of the molecule in trajectory i at time t.

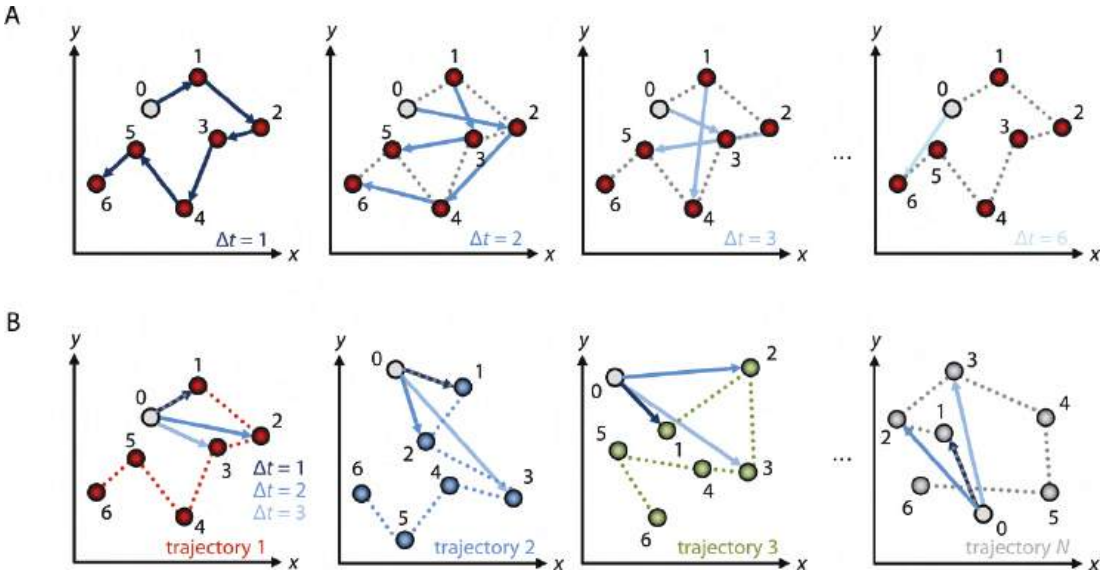


FIGURE 4.2: Determination of mean square displacements from trajectories. A: Analysis of a single trajectory. The arrows indicate the displacements for time intervals Δt (dark blue, 6 data points), $2\Delta t$ (blue, 5 data points), $3\Delta t$ (light blue, 4 data points), and $6\Delta t$ (light cyan, one data point). Displacements for the same interval Δt (depicted in the same color) are averaged, and the averages are summed according to eq. 4.20. B: Analysis of an ensemble of n trajectories, depicted in different colors (red, blue, green, gray). Arrows of the same color indicate displacements in the different trajectories for the same interval Δ. These values are averaged, and then summed according to eq. 4.21.

For two-dimensional diffusion, the mean square displacement $\langle r^2 \rangle$ follows eq. 4.22:

$$\left\langle r^2 \right\rangle\left(\tau\right) = 4D\tau \qquad \text{eq. 4.22}$$

where D is the diffusion coefficient. Thus, the diffusion coefficient can be determined from the slope of a plot of the mean square displacement $\langle r^2 \rangle$ as a function of τ (Figure 4.3). The trajectories of single molecules can be followed by imaging techniques with single-molecule resolution (Section 25.7.3).

Under certain conditions, diffusion of membrane proteins is limited to a defined area of the membrane (Box 4.2). This phenomenon is called anomalous, restricted, or *corralled diffusion* (Figure 4.3). For restricted diffusion, the mean square displacement $\langle r^2 \rangle$ is described by

$$\left\langle r^2 \right\rangle\left(\tau\right) = \frac{d^2}{3}\left(1 - \exp\left(\frac{-12D_0\tau}{d^2}\right)\right) \qquad \text{eq. 4.23}$$

where d is a measure of the dimension of the area within which the protein can undergo free diffusion, and D_0 is the diffusion coefficient at time $t = 0$. In this case, the mean square displacement reaches a limiting value that provides information on the dimension of the area within which the protein can move (Figure 4.3).

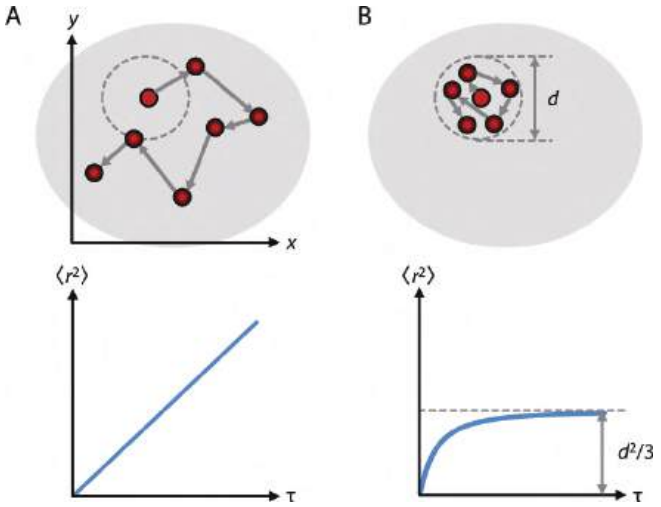


FIGURE 4.3: Free and restricted diffusion. A: A protein (red) diffuses freely within a membrane (gray). The position of the protein at different time points is indicated. The square of the distance the protein has traveled from its starting point increases linearly as a function of time. B: A protein (red) undergoes restricted (corralled) diffusion within a delimited area of the membrane (gray broken line). The mean square displacement now depends exponentially on time. The amplitude of the exponential function is related to the size of the membrane area within which the protein diffuses freely.

BOX 4.2: MEMBRANE RAFTS AND CORRALLED DIFFUSION.

The *fluid mosaic model* of lipid membranes (see Section 18.4.3.3) describes the membrane as a fluidic phase in which lipids and embedded proteins undergo free lateral diffusion. Many membrane studies are performed on reconstituted membranes with a well-defined lipid composition that differs from the composition and complexity of natural cell membranes. Diffusion coefficients of lipids and membrane proteins in reconstituted membranes are often one order of magnitude higher than diffusion coefficients within the cell membrane. The slower diffusion in biomembranes may be due to the large protein content and the molecular crowding. Furthermore, cell membranes interact with the actin cytoskeleton, and the actin filament network also affects the lateral diffusion of lipids and proteins within the membrane. In model membranes, the formation of lipid domains has been observed. These *lipid rafts* have a lipid composition that is distinct from the rest of the membrane, and are often enriched in cholesterol. The higher local fluidity within rafts may lead to corralled diffusion of transmembrane proteins, and thus to their enrichment within these membrane domains. It is still a matter of debate if these rafts also exist in natural membranes. The membrane domains forming in biological membranes are typically small (tens of nanometers) and highly dynamic. Nevertheless, rafts are thought to play an important role in regulating the local concentration of transmembrane proteins, including transporters and receptors. By spatial regulation of receptor concentration, dimerization, and activation, they might act as platforms for membrane signaling and trafficking.

4.2 THE CHEMIOSMOTIC HYPOTHESIS

The chemical potential is a driving force for chemical reactions. We have seen in Section 1.7.5 that the chemical potential depends on concentration. We now consider a system that is separated from the surroundings by a barrier that is impermeable for protons, such as a biological membrane. The chemical potential of protons inside and outside depends on the local proton concentration, and is described by

$$\mu_{H^+,\,outside} = \mu^0_{H^+} + RT \ln c_{H^+,\,outside} \qquad \text{eq. 4.24}$$

and

$$\mu_{H^+,\,inside} = \mu^0_{H^+} + RT \ln c_{H^+,\,inside} \qquad \text{eq. 4.25}$$

If the proton concentrations inside and outside are different, then

$$\mu_{H^+,\,inside} \neq \mu_{H^+,\,outside} \qquad \text{eq. 4.26}$$

The difference of the chemical potential of protons inside and outside constitutes a driving force for the equilibration of the concentrations: in the absence of the barrier, protons would spontaneously move from the compartment with higher concentration (and higher chemical potential) to the compartment with lower concentration (and lower chemical potential). The flow would continue until the concentrations in both compartments are equal, the chemical potentials are equal, and there is no driving force for further proton flux. However, the impermeable membrane prevents such an equilibration of proton concentrations, and energy can be stored in the form of a concentration gradient between protons (or other ions and molecules that cannot pass the membrane) on both sides of the membrane. The concept of storing energy in the form of concentration gradients is called the *chemiosmotic hypothesis*. The energy that is stored in a concentration gradient can drive energetically unfavorable processes. Proton gradients are a central element in the intermediate storage of energy during oxidative phosphorylation and photosynthesis (Figure 4.4). In mitochondria, energetically favorable electron transport (Section 5.6) is coupled to the transport of protons out of the mitochondria into the intermembrane space between the inner and outer membrane. In a separate reaction, protons flow back into the mitochondria through a membrane protein called F_1-F_o ATPase. The F_1-F_o ATPase couples the endergonic synthesis of ATP from ADP and phosphate to the exergonic flow of protons from outside to inside. The energy is now stored in the form of ATP, and ATP hydrolysis then provides the energy for numerous endergonic processes in the cell. Similarly, in the light reaction of photosynthesis the energy from electron transport chains in the thylakoid membrane of chloroplasts is converted into a proton gradient. The proton gradient is used to drive ATP synthesis by a transmembrane ATPase (Figure 4.4). The energy stored in a proton gradient is also called *proton-motive force* (PMF).

Biological membranes are not only barriers for protons, but also for other ions and polar molecules, which allows the establishment of ion gradients and electromotive forces. Ion gradients serve as an energy source for transport (Section 4.3). The free energy change associated with the transport of protons or other ions across the membrane into the cell is

$$\Delta G = -RT \ln \frac{c_{in}}{c_{out}} + nF\Delta\phi \qquad \text{eq. 4.27}$$

with the concentrations inside and outside of the cell, c_{in} and c_{out}, the number of charges n of the ion, the Faraday constant F, and the potential difference across the membrane $\Delta\phi$. We will discuss the membrane potential $\Delta\phi$ in Section 5.7.

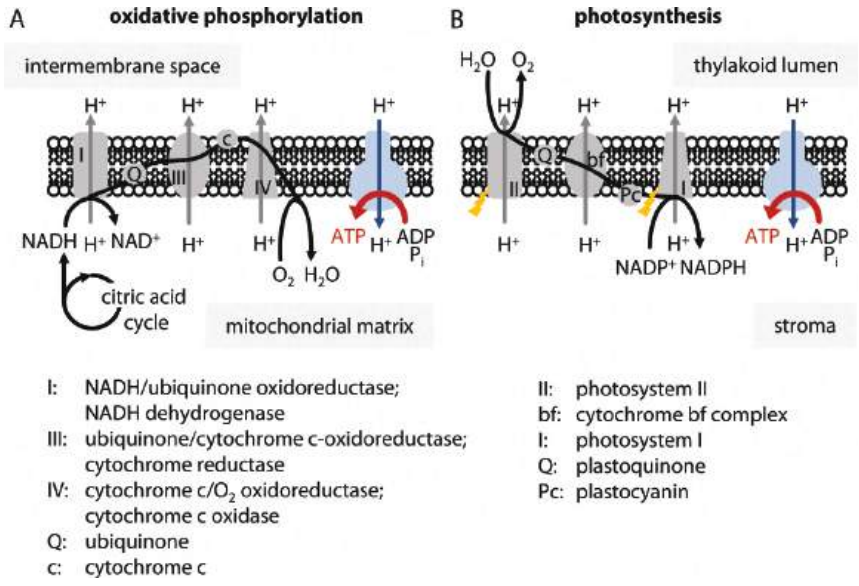


FIGURE 4.4: Proton gradients for intermediate energy storage in oxidative phosphorylation and photosynthesis. A: Oxidative phosphorylation takes place in the inner membrane of mitochondria. The citric acid cycle in the mitochondrial matrix generates reduction equivalents (NADH) that are oxidized to NAD^+ at the inner membrane by complex I. The electrons from NADH are transported through a transport chain (black line) *via* ubiquinone (Q) to complex III and *via* cytochrome c to complex IV that reduces the terminal electron acceptor oxygen to H_2O. The energetically favorable electron transport by complex I, III, and IV is coupled to transport of protons out of the mitochondrion into the intermembrane space. The proton gradient provides the energy for ATP synthesis at the ATP synthase (blue) that is coupled to influx of protons into the mitochondrion. B: The light reaction of photosynthesis takes place in the thylakoid membrane in chloroplasts. Photosystem II oxidizes H_2O to oxygen in a light-induced reaction. The electrons are transported to photosystem I *via* plastoquinone (Q), the cytochrome bf complex, and plastocyanin (Pc). Photosystem I transfers the electrons to the terminal electron acceptor in photosynthesis, $NADP^+$, and reduces it to NADPH. The energetically favorable electron transport by the cytochrome bf complex is coupled to proton transport into the thylakoid lumen. Photosystem I contributes to the increase in proton concentration in the lumen by generating two protons per oxidized H_2O molecule. The proton-motive force then drives ATP synthesis at the ATP synthase (blue) that is coupled to an efflux of protons from the thylakoid into the chloroplast stroma.

Note that the concentrations c_{in} and c_{out} in eq. 4.27 are absolute concentrations: although we have to consider concentrations relative to the standard state, the standard-state concentration c^0 is identical inside and outside (1 M), and cancels.

Flux of ions along a concentration gradient is an exergonic, spontaneous process. If the ion flux is controlled, the energy can be coupled to the endergonic synthesis of ATP (Figure 4.5). Proton gradients are exploited to drive the energetically unfavorable synthesis of ATP from ADP and phosphate. Transport of ions or molecules against a concentration gradient, and thus building up concentration gradients, is an endergonic process, and requires energy input. Concentration gradients can be established by coupling transport to exergonic reactions, such as ATP hydrolysis, or redox reactions (oxidative phosphorylation, photosynthesis). Uncoupling flux along concentration gradients from ATP synthesis can be used to generate heat without muscle activity (Box 4.3). The reaction can also be reversed, and ATP hydrolysis can provide the energy for establishing proton (or other) gradients (Figure 4.5). The ATP-dependent establishment of an ion concentration gradient is important for the generation and maintenance of membrane potentials (Section 5.7).

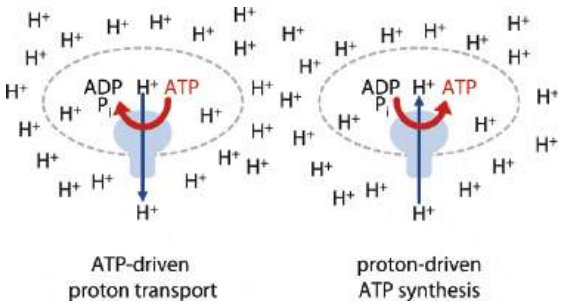


FIGURE 4.5: Concentration gradients and energy storage. The energy of ATP hydrolysis can be used to drive proton transport against a concentration gradient, converting the energy into a proton gradient (left). Conversely, the energy stored in a proton gradient can drive the synthesis of ATP from ADP and phosphate, coupled to the flux of protons along the concentration gradient (right).

BOX 4.3: UNCOUPLING THE PROTON-MOTIVE FORCE FROM ATP SYNTHESIS.

Instead of coupling the proton-motive force to ATP synthesis, the energy can also be converted into heat by uncoupled influx of protons. This uncoupling is possible in the mitochondria of brown fat tissue. Instead of passing through the proton channel that is coupled to the ATP synthase, protons re-enter the cell through the ligand-gated proton channel thermogenin (see Section 4.3). Thermogenin is closed in the unliganded state, but opens in response to binding of fatty acids. The fatty acids are generated in a signaling cascade. Upon binding of noradrenalin to receptors on the cell membrane of brown fat cells, a cAMP signaling cascade leads to activation of a lipase that hydrolyzes triglycerides to glycerol and fatty acids. The fatty acids are the activating ligands for thermogenin and lead to opening of the channel. Protons can now directly enter the mitochondria, and the energy of the proton gradient is not stored in the form of ATP, but directly released as heat. This process is used by babies, and also by hibernating mammals. Uncoupling allows the generation of heat without muscle activity in the form of trembling. Membrane-permeable hydrophobic molecules that are weak acids, such as 2,4-dinitrophenol, can also uncouple the proton-motive force from ATP synthesis by diffusing through the lipid membrane and transporting protons. While these substances are not very healthy, they have been useful tools in biochemical studies of the oxidative phosphorylation pathway.

4.3 ACTIVE AND PASSIVE TRANSPORT

We have seen before that the flux of ions or molecules from high concentration to low concentration along a concentration gradient is exergonic and thus in principle spontaneous. However, biological membranes act as barriers and impede spontaneous flow. Pore-forming proteins spanning the membrane can provide a channel for molecules to pass the membrane (Figure 4.6). Flow of molecules through these pores along a concentration gradient is called *passive transport*. Passive transport is typically regulated, and opening of the channel is triggered by ligand binding (*ligand-gated channel*) or by a change in membrane potential (*voltage-gated channel*) (Figure 4.6).

Transport of ions or molecules against a concentration gradient, from low to high concentrations, is an endergonic process: it does not occur spontaneously, but requires energy. Active transporters use the energy of ATP hydrolysis (*primary active transport*) or the energy of concentration gradients of other molecules (*secondary active transport*) to drive the transport of their cargo. In

secondary active transport, molecules co-transported with the cargo can flow in the same (symport) or the opposite direction (antiport) (Figure 4.6).

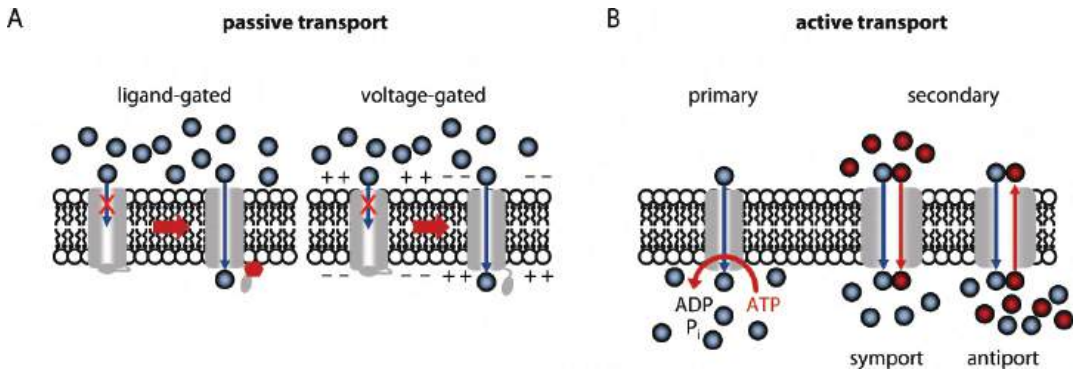


FIGURE 4.6: Transport across biological membranes. A: Passive transport along concentration gradients is mediated by channels. Channel opening is typically regulated, either by binding of ligands (ligand-gated channel) or by changes in the membrane potential (voltage-gated channel). Active transport against concentration gradients requires energy. In primary active transport, this energy stems from ATP hydrolysis. In secondary active transport, the energy is provided by the electromotive force of a second particle (red) that is co-transported along its own concentration gradient. Co-transport of this particle can occur in the same direction as the cargo (symport) or in the opposite direction (antiport).

The transport cycle of a primary active transporter needs to couple ATP hydrolysis to transport of the cargo. This coupling can be achieved by cycling of the transporter through four different states (Figure 4.7). At the beginning of the transport cycle, the cargo binding site of the transporter is exposed to the left-hand side of the membrane where the cargo will be taken up. The energy barrier for binding is low, and the cargo is rapidly bound (see Section 15.2 for the relation between energy barriers and reaction velocities). The bound state is a global minimum in the energy profile, and the cargo is stably bound. The energy barrier to the other side of the membrane is high, however, making release of the cargo to the right very slow. The transporter then switches into a second state with its binding site exposed toward the other side of the membrane. This transition can be induced *e.g.* by ATP-dependent phosphorylation. In this second state, the energy profile of the transporter-cargo interaction is altered: the bound state is now only a local minimum in the energy profile. The energetic barrier toward the left is high, and release toward the left is slow. Release to the right, however, is now favored both thermodynamically (energetically "downhill", exergonic) and kinetically (low-energy barrier), and the cargo will be rapidly released to the right. At this stage one ATP has been hydrolyzed, and one cargo molecule has been transported from left to right. For further transport cycles, the transporter needs to be regenerated. Cleavage of the phosphate group will convert the transporter into its original conformation. The high-affinity cargo binding site is exposed toward the left, and the transporter is ready for another round of ATP-driven transport. Some transporters use the energy of light instead of ATP to transport ions against concentration gradient. In these cases, absorption of light and relaxation from the excited state (Sections 21.1 and 21.2) triggers conformational switches of the transporter. We will revisit these light-driven ion pumps when we discuss the absorption of light and optical spectroscopy (Section 21.2).

The ATP-driven transport of protons against a concentration gradient is not too different from the reverse Carnot cycle (Section 1.3.5.1): ATP hydrolysis provides the energy (w) to take up material from the low concentration side (heat q from the colder reservoir) and transport it to the high concentration side (heat q released to the warmer reservoir). It also shows resemblance to reverse osmosis, where external pressure serves as the driving force for the non-spontaneous movement of solvent molecules against a concentration gradient (Section 1.7.7).

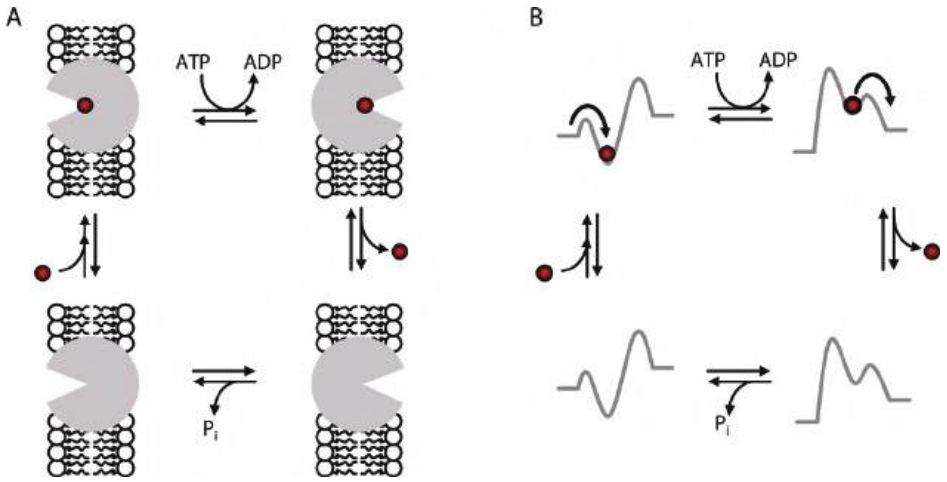


FIGURE 4.7: Four-state model of a transporter: a thermodynamic cycle with energy input. A: A transporter exists in two conformational states, one state with the high-affinity binding site for the cargo facing the exterior of the cell (left), and a second state with the low-affinity binding site facing toward the interior (right). Switching between these two conformations is achieved by phosphorylation and dephosphorylation. In both states, the transporter can be empty, or bound to cargo. For 100% coupling of ATP hydrolysis (phosphorylation) to transport of cargo, the transporter has to spend sufficient time in the outward facing state to bind cargo before the conformation is switched by phosphorylation. Similarly, it has to spend sufficient time in the inward facing state to ensure that the cargo is released before dephosphorylation switches it back to the outward facing state. B: Energy diagrams for the four states.

If the conformational switch of the transporter only occurs upon phosphorylation and dephosphorylation, and binding only occurs from the left and release only occurs to the right, then ATP hydrolysis and cargo transport are perfectly coupled; the coupling efficiency is 100%. To ensure high coupling efficiencies, timing has to be optimized: first, the transporter must stay long enough in the state with the high-affinity binding site toward the left, leaving enough time for the cargo to bind. If the transporter is phosphorylated before the cargo has bound, the transporter will undergo a futile cycle: one ATP will be hydrolyzed without a transport event. Second, the reporter needs to stay in the phosphorylated state, with the low-affinity binding site exposed to the right, long enough for the cargo to dissociate. Dephosphorylation of the transporter before the cargo has dissociated will bring the cargo back to the left, without net transport. The ATP has already been hydrolyzed at this stage, and premature switching back thus also leads to a futile cycle. The rates of cargo binding and release therefore have to be matched with the conformational cycle that is regulated by phosphorylation/dephosphorylation. To ensure that no cargo is transported in the absence of ATP hydrolysis, the conformational switch needs to be coupled strictly to phosphorylation and dephosphorylation. Spontaneous switching of a cargo-bound transporter without phosphorylation will lead to transport of one cargo molecule without ATP hydrolysis, which is referred to as *leakage*.

4.4 DIRECTED MOVEMENT BY THE BROWNIAN RATCHET

We have seen in Section 4.1 that thermal motion of molecules is a random process, and that molecules diffuse in no preferred direction in the absence of concentration gradients. However, it is possible to bias Brownian motion under certain conditions, and to enforce directional movement. We start with a thought experiment, and imagine a number of electric dipoles that are arranged head-to-tail to form a track of alternating charges (Figure 4.8). A charged particle above this track experiences repulsive forces near a charge of the same sign, and attractive electrostatic forces when it is located next to an opposite charge. We can translate this behavior into an energy landscape. For a negatively charged particle, the potential energy is low in the vicinity of positive charges on the track (attraction), and high close to negative charges (repulsion). Overall the particle therefore experiences a *saw-tooth-shaped energy potential* with periodic minima and maxima along the

track (Figure 4.8). The distance between minima and maxima corresponds to the distances of the charges on the track. In thermal equilibrium, the preferred position of the particle along the track is determined by the energetic minima or wells in the saw-tooth potential. The thermal energy of the particles enables small excursions left and right, up to a potential energy level that equals their thermal energy. If the thermal energy is sufficient, the particle can cross the energy barrier to the left or to the right, and reach the next energetic minimum to the left or right along the track. The probability for such transitions depends on the height of the energy barrier (see Section 15.2), which is identical for movement in either direction. In other words, crossing to the next well left or right is of equal probability, and the particle cannot undergo directed movement under these conditions.

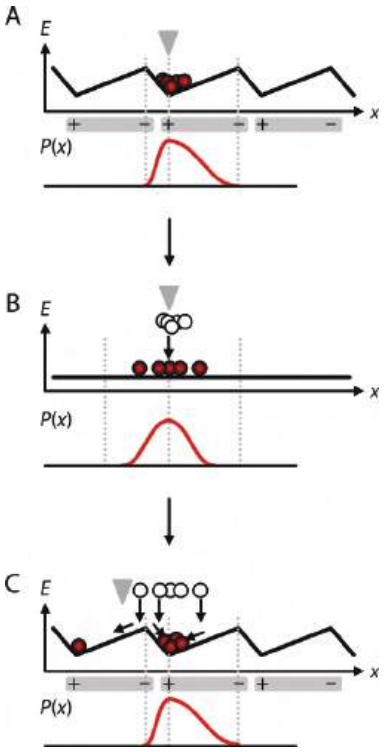


FIGURE 4.8: Principle of a Brownian ratchet. A: Negatively charged particles above dipoles aligned head to tail experience a saw-tooth energy profile: their potential energy is low at the positive end of the dipoles (energetic minimum) and high energy at the negative end (energy maximum). In the presence of the saw-tooth potential, the particles distribute around the energetic minimum. Their probability distribution $P(x)$ (red) is a function of the position x and is asymmetric, reflecting the asymmetry of the potential. As the barrier is equally high to either side, the probability for a particle to cross over to the left or right well is equal. B: If the saw-tooth profile is switched off (*e.g.* by neutralizing the charge of the particle in a chemical reaction), the energy profile is flat, and the potential energy is position-independent. The particles can now freely diffuse to left and right. The probability to move by a certain distance to left and right by diffusion is equal, the probability distribution of the particles at the end (red) is symmetric to the starting position. C: When the asymmetric saw-tooth profile is switched back on, the particles again experience a potential that depends on their actual position and will equilibrate to the position of the nearest energetic minimum. Because of the asymmetry of the potential, the position of the next barrier to the left is more likely to be crossed than that of the next barrier to the right. In our example, one particle is beyond the next barrier to the left at the end of the off-period, and equilibrates into the neighboring well on the left, while the remaining particles end up in the original well. The gray triangle indicates the mean position of the particles. The mean position has moved to the left in this one cycle.

Now imagine we switch off the electric potential. One way to achieve this would be by neutralizing the charge of the particle, *e.g.* by protonation of a negatively charged particle. The uncharged particle does not interact with the charges on the track. Therefore, the potential energy of the particle no longer depends on its position (Figure 4.8). The uncharged particle undergoes Brownian motion

starting from the equilibrium position it occupied while the saw-tooth potential was still present. With a flat potential, the probabilities for excursions to left and right are equal, or more specifically, the probability for an excursion by Δx to the left is equal to the probability for the same excursion by Δx to the right. As a consequence, the probability to find the particle at a certain position along the track now symmetrically decreases from its starting position. Up to this point, there is still no directionality in the system.

What happens if we switch the potential back on by again conferring a negative charge to the particle, *e.g.* by deprotonation? The particle experiences the same saw-tooth potential as in the beginning. Its location is again determined by the energy landscape, and it relaxes to the next energetic minimum that it can reach in an energetically favorable reaction (energetically "downhill"). At this point of the thought experiment it is important to take note of the asymmetry of the saw-tooth energy potential (Figure 4.8): the change in potential along the x-axis is much steeper to the left of the minimum than to the right. As a consequence, the particle has to diffuse only a short distance Δx during the time the potential was "off" to end up in a place corresponding to the next well to the left. To reach the next well to the right requires that the particle travels a much longer way. Traveling a short distance in a fixed time window is more probable than traveling a longer distance. Correspondingly, the chances that we catch the particle in a position that leads it into the next left well when we switch on the potential are higher than the chances that the particle has traveled far enough to be in the position of the next right well. Hence, the probability for a movement of the particle to the left is higher than for a movement to the right (Figure 4.8). By switching the saw-tooth energy potential on and off, we have biased the Brownian motion of the particle to directed movement to the left! Using the same principle, we can force a set of particles to undergo an overall movement to the left by switching the potential on and off, and generate directed motion. Such a rudimentary motor is called a *Brownian motor* or *Brownian ratchet*.

To achieve directed movement, the timing of the changes in potential is important. The potential needs to be off long enough for the particles to cover sufficient distance to leave the realm of one energy well, and to end up in the left neighboring well once the potential is switched on again. Similarly, the potential needs to be on long enough for the particles to equilibrate in the position of the energetic minimum. In our example, the timing is determined by the time point of deprotonation (to generate the negatively charged particle and establish the saw-tooth potential) and the time point of protonation (to generate an uncharged particle and switch off the saw-tooth potential). More generally speaking, the kinetics of the chemical reaction that modulates the interaction of the particle with the track is crucial for efficient directional motion. Kinetics and energetics need to be balanced for efficient movement, however. The asymmetry of the saw-tooth potential determines how strongly movement is biased into one direction (Figure 4.9). The distance between the minimum and the next maximum in the potential determines the distance the particle has to travel to cross over to the next well. The shorter this distance, the shorter the time the particle needs to cover this distance, and the more probable it is for the particle to reach this well. A Brownian ratchet with a high spatial repetition frequency of the periodic potential can be driven by a fast chemical reaction, whereas a ratchet with a lower repetition frequency must be driven by a slower reaction. The shape of the potential and the distances between minima and maxima are determined by the structure of the units that assemble linearly into the track.

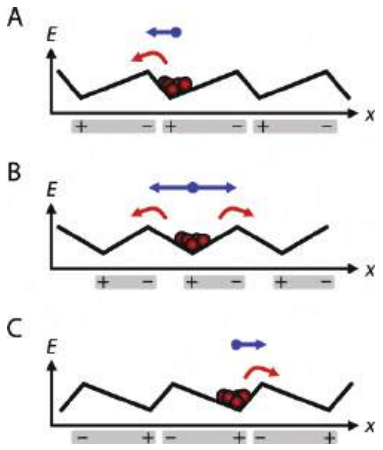


FIGURE 4.9: Shape of the potential and direction of movement. Energy potential for negatively charged particles above a track made of electric dipoles. The traveling distance to reach the neighboring well is indicated by the blue arrow, the preferred direction of movement by the red arrow. A: Asymmetric saw-tooth potential that biases motion to the left. B: Symmetric saw-tooth potential. Movement to the left and to the right is equally probable. C: Asymmetric saw-tooth potential that biases motion to the right.

The remaining question from a thermodynamic point of view is "where does the energy for the directed movement come from?" In our model, switching the potential on and off requires energy, and this is the energy that is converted into directional movement. The energy potential was switched on and off because of the protonation/deprotonation reaction that neutralized or generated the negative charge of the particle. In general terms, the energy is thus provided by the chemical reaction that modulates the interaction between particle and track. It is important to note that a Brownian ratchet can be driven only by a reaction that is not in equilibrium (see Box 1.7). This is quite clear, because ΔG for a reaction in equilibrium is zero (Section 2.1; eq. 2.7), and this reaction cannot perform work.

Directional movement in biology is often achieved by biasing Brownian motion. The energy for this movement is typically provided by ATP hydrolysis. The proton transporter we discussed in Section 4.3 is an example of a Brownian ratchet: the transporter is switched between two conformations in an ATP-dependent process (phosphorylation/dephosphorylation). The phosphorylation/dephosphorylation reaction alters the interaction of the transporter with the cargo, and modulates the interaction energy landscape. As a result, one conformation of the transporter and the resulting energy potential enable rapid cargo binding from the outside with high affinity, and the other conformation is characterized by a low cargo affinity and rapid cargo dissociation to the inside. Conceptually the transporter therefore corresponds to the particle in our model Brownian ratchet, and the cargo corresponds to the track. The timing of the chemical reaction is key for the movement of the cargo across the membrane.

A motor example is the movement of myosin along actin filaments. The periodic structure of actin filaments (the track) generates a periodic energy potential that governs the interaction of myosin (the particle) with actin. Myosin is an ATPase: it binds and hydrolyzes ATP, and thereby alternates between an ATP and an ADP state that interact differently with the actin filament. Depending on its nucleotide state, myosin either binds tightly to the actin filament (potential on) or it binds with low affinity and becomes detached from the filament (potential off). The ATP- and ADP-bound states of myosin thus represent the charged and uncharged state of the particle in our conceptual model. The rate constants for ATP binding, hydrolysis, and dissociation of the products control how long each nucleotide state of myosin persists, and at which point in time the potential is switched on and off. The structure of actin and the periodicity of the actin filament determine the shape and periodicity of the interaction potential. Of course, the interactions between myosin and actin are much more complex than simple on- and off-switching of electrostatic interactions. Nevertheless, the principle of the movement of myosin relative to actin and of other biological motors relative to their track can be described by the conceptually simple model for a Brownian ratchet.

In summary, the Brownian ratchet describes the behavior of molecular motors on the basis of interactions between a moving unit and its track, modulated by a chemical reaction that provides energy input and the timing of changes in the interaction potential. We have restricted our considerations to electrostatic interactions, but biological motors can realize much more complex interactions with the track. Conformational dynamics subtly modulate these interactions, and influence the timing of changes, leading to higher levels of complexity and offering possibilities for regulation.

QUESTIONS

4.1 Explain diffusion along concentration gradients based on the chemical potential and on entropy considerations.

4.2 A protein has a diffusion coefficient in the cell of $D = 10^{-11}$ m^2 s^{-1}. How long does it take the molecule to traverse the cell (diameter $r = 1$ µm)?

4.3 The concentration of chloride ions is higher in blood (0.1 M) than in brain cells (0.05 M). What is ΔG for the transport of Cl^- from the blood into the brain and from the brain into the blood at 37°C?

4.4 The interior of our stomach has a pH of 2. The cells in the lining mucosa have a neutral pH (pH 7.2). Calculate the energy that is required for the secretion of H^+ from the mucosa cells into the gastric lumen.

4.5 The change in free energy for ATP hydrolysis is $\Delta G^{0'} = -31$ kJ mol^{-1}. What is the minimal proton gradient that is required for ATP synthesis under standard conditions and at intracellular concentrations of $[phosphate] = 10^{-2}$ M, $[ADP] = 10^{-4}$ M, and $[ATP] = 10^{-3}$ M? Calculate the required pH outside the cell assuming a neutral pH in the cytoplasm. $T = 25$°C.

4.6 The movement of two fluorescently labeled membrane receptors R1 and R2 within the membrane is imaged by total internal reflection microscopy. The following positions are determined as a function of time:

t (s)	R1 (x/y)	R2 (x,y)
0	1.2/2.8	4.6/9.9
1	3.3/2.1	5.2/8.8
2	4.6/2.7	3.8/9.2
3	5.4/6.8	4.8/8.3
4	6.0/4.8	5.2/9.4
5	3.2/6.2	3.9/8.4

What can you conclude about the types of diffusion and the diffusion coefficients of R1 and R2?

4.7 How can the movement of polymerases along DNA be described by the Brownian ratchet model?

REFERENCES

Astumian, R. D. and Derenyi, I. (1998) Fluctuation driven transport and models of molecular motors and pumps. *Eur Biophys J* **27**(5): 474–489.
· review on the principle of a Brownian ratchet

Ait-Haddou, R. and Herzog, W. (2003) Brownian ratchet models of molecular motors. *Cell Biochem Biophys* **38**(2): 191–214.
· review on the Brownian ratchet and its application to biological motors

Astumian, R. D. (2005) Biasing the random walk of a molecular motor. *Journal of Physics: Condensed Matter* **17**(47): S3753.
· Brownian ratchet model and application to ATP-dependent enzymes and the movement of kinesin along microtubules

Engelman, D. M. (2005) Membranes are more mosaic than fluid. *Nature* **438**(7068): 578–580.
· introductory review article on the mosaicity of membranes

Ernst, D. and Kohler, J. (2013) Measuring a diffusion coefficient by single-particle tracking: Statistical analysis of experimental mean squared displacement curves. *PCCP* **15**(3): 845–849.
· treatment of the effect of the number of data points for the analysis of trajectories

Heinemann, F., Vogel, S. K. and Schwille, P. (2013) Lateral membrane diffusion modulated by a minimal actin cortex. *Biophys J* **104**(7): 1465–1475.
· original paper on the effect of the actin cytoskeleton on protein diffusion in membranes

Kusumi, A., Nakada, C., Ritchie, K., Murase, K., Suzuki, K., …, Fujiwara, T. (2005) Paradigm shift of the plasma membrane concept from the two-dimensional continuum fluid to the partitioned fluid: High-speed single-molecule tracking of membrane molecules. *Annu Rev Biophys Biomol Struct* **34**: 351–378.
· review on current models for the diffusion of proteins in membranes

Lingwood, D. and Simons, K. (2010) Lipid rafts as a membrane-organizing principle. *Science* **327**(5961): 46–50.
· review on rafts as platforms and their dynamics in cellular membranes

Mitchell, P. (2011) Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. *Biochim Biophys Acta* **1807**(12): 1507–1538.
· reprint of the original paper from Peter Mitchell on the chemiosmotic hypothesis

Sanderson, J. M. (2012) Resolving the kinetics of lipid, protein and peptide diffusion in membranes. *Mol Membr Biol* **29**(5): 118–143.
· review on diffusion of lipids and proteins in membranes

Simons, K. and Gerl, M. J. (2010) Revitalizing membrane rafts: New tools and insights. *Nat Rev Mol Cell Biol* **11**(10): 688–699.
· review article on the role of membrane domains in signaling and trafficking and insight from modern fluorescent microscopy methods

Singer, S. J. and Nicolson, G. L. (1972) The fluid mosaic model of the structure of cell membranes. *Science* **175**(4023): 720–731.
· original paper that introduces the fluid mosaic model of membranes



5 Electrochemistry

Electrochemistry comprises reactions that involve charged particles, and the transfer of electrons. For their thermodynamic description, we have to expand our definition of the chemical potential, and we have to include a term that represents the electrochemical work associated with the transport of charges in an electric field. We define the *electrochemical potential* $\tilde{\mu}_j$ of ions in the presence of electric fields as (eq. 5.1):

$$\tilde{\mu}_j = \mu_j + z_j F \phi \qquad \text{eq. 5.1}$$

μ_j is the chemical potential, z is the number of charges of the ion, F is the Faraday constant (96500 C mol^{-1}, *i.e.* the charge of one mol of elementary charges: the Faraday constant is the product of the elementary charge of one electron and the Avogadro constant), and ϕ is the electrostatic potential. The electrochemical potential is the driving force for electrochemical reactions, in the same way that the chemical potential is the driving force for a chemical reaction in the absence of electrochemical work. For uncharged particles ($z = 0$) or in the absence of an electrostatic potential, eq. 5.1 simplifies to the general definition of the chemical potential.

5.1 REDOX REACTIONS AND ELECTROCHEMICAL CELLS

Reactions that involve the transfer of electrons are called *redox reactions*. A *reduction* reaction is the uptake of one (or more) electrons. The reverse reaction, the release of one or more electrons is called an *oxidation*. As electrons are normally not stable in solution, reduction and oxidation are usually coupled, leading to the term *redox reaction*. A physiologically important oxidation is the reaction of NADH/H$^+$ to NAD$^+$:

$$\mathrm{NADH/H^+ \longrightarrow NAD^+ + 2\,e^- + 2\,H^+} \qquad \text{scheme 5.1}$$

A reduction that is central for our lives is the conversion of molecular oxygen to water:

$$\mathrm{O_2 + 4\,e^- + 4\,H^+ \longrightarrow 2\,H_2O} \qquad \text{scheme 5.2}$$

The overall redox reaction is

$$2\,NADH/H^+ + O_2 \longrightarrow 2\,NAD^+ + 2\,H_2O \qquad \text{scheme 5.3}$$

We have encountered these reactions in the context of oxidative phosphorylation (Figure 4.4). To understand the quantitative concepts governing redox reactions, we will start with a very simple inorganic reaction in solution. If we add elemental zinc to a solution of a copper(II) salt, *i.e.* containing Cu^{2+} ions, the following reaction will take place:

$$Zn + Cu^{2+} \longrightarrow Zn^{2+} + Cu \qquad \text{scheme 5.4}$$

The zinc metal will dissolve and metallic copper will appear. Conceptionally and physically, we can separate this reaction into two half-reactions that occur in separate *half-cells* (Figure 5.1). Imagine we have a zinc electrode in a solution of zinc ions. At the electrode, zinc and zinc ions are in equilibrium. Zinc can release two electrons and will be converted into a zinc ion. Inversely, zinc ions can take up two electrons and will be deposited on the electrode as elemental zinc. This is also a redox reaction, in which the oxidant and the reducing agent are the same element.

$$Zn^{2+} + 2\,e^- \rightleftharpoons Zn \qquad \text{scheme 5.5}$$

On the other hand, we can immerse a copper electrode in a solution containing copper ions, and will have the corresponding equilibrium between copper and copper ions.

$$Cu^{2+} + 2\,e^- \rightleftharpoons Cu \qquad \text{scheme 5.6}$$

For zinc, the equilibrium is on the side of the ion, meaning the release of two electrons is an exergonic reaction, and thus the dissolution of zinc metal is the spontaneous direction: the zinc from the electrode is oxidized. For copper, on the other hand, the equilibrium is on the side of elemental copper. Copper ions from the solution take up electrons, and they are reduced to copper that is deposited on the electrode.

We can now combine the two half-cells to an *electrochemical cell* (Figure 5.1). Due to the different positions of the two ionization equilibria, the electrochemical potential of the two electrodes will be different, and this difference of potential leads to a (measurable) voltage. If we connect the half-cells with a conducting material, electrons released by zinc oxidation at the zinc electrode can then flow to the copper electrode, and can be used to reduce copper ions in the other half-cell: reduction and oxidation are coupled in a *redox reaction*.

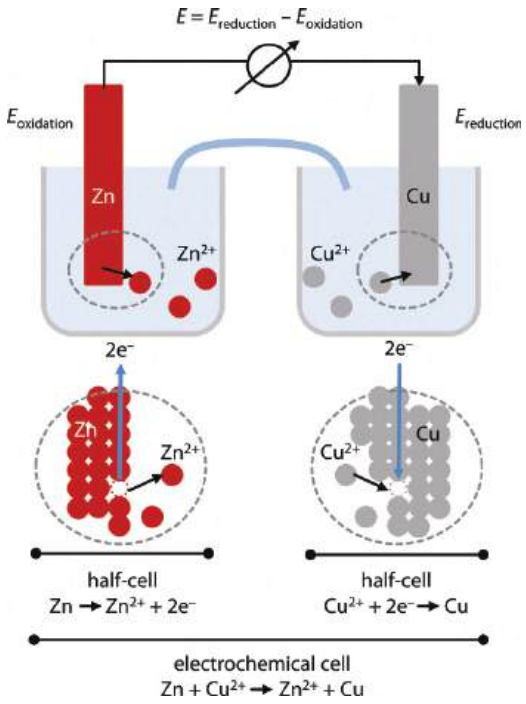


FIGURE 5.1: Electrochemical half-cells and cell. When we place zinc metal (electrode) into a solution of a zinc salt, an equilibrium between zinc ions (Zn^{2+}) and elemental zinc (Zn) is established, leading to an electrochemical potential of the zinc half-cell (left). Similarly, when we place copper metal (Cu) into a solution of a copper salt, an equilibrium between Cu and Cu^{2+} ions is established, causing an electrode potential. For Zn/Zn^{2+}, the equilibrium is on the side of the ions. Zn will be oxidized to Zn^{2+}, and the surplus of electrons causes a negative electrode potential. For Cu/Cu^{2+}, the equilibrium is on the side of Cu, and the Cu^{2+} ions will be reduced at the electrode, leading to the deposition of Cu and a depletion of electrons (positive electrode potential, right). We can combine the two half-cells to an electrochemical cell, in which the two solutions are connected by an ion bridge. The difference in electrode potentials $E_{oxidation}$ and $E_{reduction}$ leads to a measurable voltage between the two half-cells, the electromotive force (see eq. 5.17).

The free energy change associated with the redox reaction (Scheme 5.4) can be calculated from the electrochemical potentials according to eq. 5.2:

$$\Delta G = \tilde{\mu}_{Zn^{2+}} + \tilde{\mu}_{Cu} - \tilde{\mu}_{Cu^{2+}} - \tilde{\mu}_{Zn} \qquad \text{eq. 5.2}$$

The electrochemical potentials of the (solid) electrode materials equal the standard chemical potentials (Section 1.7.5). If the electrochemical cell operates at standard pressure of 10^5 Pa (1 bar), then

$$\tilde{\mu}_{Cu} = \tilde{\mu}_{Zn} = 0 \qquad \text{eq. 5.3}$$

(Section 1.7.3). With the help of eq. 1.162 that expresses the concentration dependence of the chemical potential we can then re-write eq. 5.2 as

$$\Delta G = \tilde{\mu}^0_{Zn^{2+}} - \tilde{\mu}^0_{Cu^{2+}} + RT \ln \frac{\left[Zn^{2+}\right]}{\left[Cu^{2+}\right]} = \Delta G^0 + RT \ln \frac{\left[Zn^{2+}\right]}{\left[Cu^{2+}\right]} \qquad \text{eq. 5.4}$$

The free energy change of the redox reaction equals the electrical work, which is the product of the transported charge (per mol) and the potential difference between the electrodes, $\Delta\phi$, in V. The transported charge, in turn, can be calculated from the Faraday constant F (96500 C mol^{-1}) and the number of electrons n that are exchanged:

$$\Delta G = -w_{el} = -n \cdot F \cdot \Delta\phi \qquad \text{eq. 5.5}$$

The negative sign in eq. 5.5 takes into account that the spontaneous reaction of the electrochemical cell with a negative ΔG allows the cell to perform work on the surroundings, and is in accordance with the system-centered perspective (Section 1.1). We name the potential difference of an electrochemical cell working under reversible conditions the *electromotive force E* (or EMF, measured in V), and obtain eq. 5.6:

$$\Delta G = -n \cdot F \cdot E \qquad \text{eq. 5.6}$$

Under standard conditions, eq. 5.6 becomes

$$\Delta G^0 = -n \cdot F \cdot E^0 \qquad \text{eq. 5.7}$$

with the *standard electromotive force* E^0 of the electrochemical cell. By combining eq. 5.6 and eq. 5.7 with eq. 5.4, we can write

$$E = E^0 - \frac{RT}{nF} \ln \frac{\left[Zn^{2+}\right]}{\left[Cu^{2+}\right]} \qquad \text{eq. 5.8}$$

Eq. 5.8 allows us to calculate the electromotive force of an electrochemical cell from the standard electromotive force and the actual concentrations of the ions involved.

5.2 TYPES OF HALF-CELLS

There are different types of half-cells. So far, we have considered half-cells consisting of a metal electrode immersed in a solution of a salt of this metal. Half-cells can also consist of a gas and an ionic solution. A prominent example we will come back to when we discuss reference measurements is the hydrogen electrode, where molecular hydrogen (H_2) dissociates into H atoms at a platinum electrode, and the H atoms are in equilibrium with hydronium ions (hydrated H^+):

$$2\,H^+ + 2\,e^- \rightleftharpoons H_2 \qquad \text{scheme 5.7}$$

A half-cell is also obtained when the same species is present in different oxidation states, such as

$$Fe^{3+} + e^- \rightleftharpoons Fe^{2+} \qquad \text{scheme 5.8}$$

or

$$Cu^{2+} + e^- \rightleftharpoons Cu^{+} \qquad \text{scheme 5.9}$$

Fe^{2+}/Fe^{3+} and Cu^{2+}/Cu^{+} are redox pairs in components of the electron transport chains in oxidative phosphorylation and photosynthesis (see Section 5.6).

The last type of half-cells consists of a metal electrode, covered with a layer of an insoluble salt of the same metal and in contact with a solution that contains the anion of this salt. Examples are the silver electrode:

$$AgCl\,(s) + e^- \rightleftharpoons Ag\,(s) + Cl^- \qquad \text{scheme 5.10}$$

or the calomel electrode

$$Hg_2Cl_2\,(s) + 2\,e^- \rightleftharpoons 2\,Hg\,(s) + 2\,Cl^- \qquad \text{scheme 5.11}$$

Silver and the calomel electrodes are frequently used as reference half-cells in pH measurements (Section 5.5).

5.3 STANDARD ELECTRODE POTENTIALS

It is not possible to measure the small potential differences between the solution and the electrode of a half-cell directly. Therefore, only relative half-cell electrode potentials against a second half-cell can be measured. The hydrogen electrode is used as a reference half-cell in these measurements, and its standard electrode potential has been defined as $E^0 = 0$ V. According to the standard electrode potentials, measured against the reference hydrogen electrode, different half-cells are ordered in the *electrochemical series* (Table 5.1), from highest positive to the highest negative electrode potential. In a half-cell with a negative standard potential, the release of electrons, the oxidation, is favored. In a half-cell with a positive standard potential, the uptake of electrons, the reduction, is favored.

TABLE 5.1:
The electrochemical series (I). Selection of standard electrode potentials in electrochemical order.

reduction (half-)reaction	E^0 (V)
$AgCl + e^- \rightarrow Ag + Cl^-$	+1.98
$Cl_2 + 2\ e^- \rightarrow 2\ Cl^-$	+1.36
$Ag^+ + e^- \rightarrow Ag$	+0.80
$Cu^{2+} + 2\ e^- \rightarrow Cu$	+0.34
$2\ H^+ + 2\ e^- \rightarrow H_2$	0.00
$Zn^{2+} + 2\ e^- \rightarrow Zn$	–0.76
$Na^+ + e^- \rightarrow Na$	–2.71

Standard potentials for biochemical redox systems are often given in the reference system of the biochemical standard state at pH 7, and are marked as $E^{0\prime}$. Biochemical standard potentials for biologically relevant reduction reactions are summarized in Table 5.2.

TABLE 5.2:
The electrochemical series (II). Biochemical standard potentials of biologically relevant reduction reactions in electrochemical order.

reduction (half-)reaction	$E^{0\prime}$ (V)
$O_2 + 4\ H^+ + 4\ e^- \rightarrow 2\ H_2O$	+0.81
$NO_3^- + 2\ H^+ + 2\ e^- \rightarrow NO_2^- + H_2O$	+0.42
$Fe^{3+} + e^- \rightarrow Fe^{2+}$ (cytochrome f)	+0.36
$Cu^{2+} + e^- \rightarrow Cu^+$ (plastocyanin)	+0.35
$Cu^{2+} + e^- \rightarrow Cu^+$ (azurin)	+0.30
$O_2 + 2\ H^+ + 2\ e^- \rightarrow H_2O_2$	+0.30
$Fe^{3+} + e^- \rightarrow Fe^{2+}$ (cytochrome c)	+0.25
$Fe^{3+} + e^- \rightarrow Fe^{2+}$ (cytochrome b)	+0.08
dehydroascorbic acid + $2\ H^+ + 2\ e^-$ → ascorbic acid	+0.08
coenzyme Q + $2\ H^+ + 2\ e^-$ → coenzyme QH_2	+0.04
fumarate + $2\ H^+ + 2\ e^-$ → succinate	+0.03
vitamin K_1 (ox) + $2\ H^+ + 2\ e^-$ → vitamin K_1 (red)	–0.05
oxaloacetate + $2\ H^+ + 2\ e^-$ → malate	–0.17
pyruvate + $2\ H^+ + 2\ e^-$ → lactate	–0.18
acetaldehyde + $2\ H^+ + 2\ e^-$ → ethanol	–0.20
riboflavin (ox) + $2\ H^+ + 2\ e^-$ → riboflavin (red)	–0.21
FAD + $2\ H^+ + 2\ e^-$ → $FADH_2$	–0.22
glutathione (ox) + $2\ H^+ + 2\ e^-$ → glutathione (red)	–0.23
lipoic acid (ox) + $2\ H^+ + 2\ e^-$ → lipoic acid (red)	–0.29
$NAD^+ + H^+ + 2\ e^-$ → NADH	–0.32
cystine + $2\ H^+ + 2\ e^-$ → 2 cysteine	–0.34
acetyl-CoA + $2\ H^+ + 2\ e^-$ → acetaldehyde + CoA	–0.41
$2\ H_2O + 2\ e^- \rightarrow H_2 + 2\ OH^-$	–0.42
ferredoxin (ox) + e^- → ferredoxin (red)	–0.43
$O_2 + e^- \rightarrow O_2^-$	–0.45

Note that a negative value for E^0 corresponds to a positive ΔG^0, meaning the reduction reaction is thermodynamically disfavored, and the oxidation reaction is thermodynamically favored. In contrast, a positive value for E^0 corresponds to a negative ΔG^0 for the reduction reaction, *i.e.* the

reduction is thermodynamically favored. It is important to note that tabulated values for standard electrode potentials always refer to the reduction reaction. If the reaction proceeds in the opposite direction, toward oxidation, the sign of the standard electrode potential has to be inverted for calculations of the associated free energy change. We will come back to this sign inversion in Section 5.4.

5.4 THE NERNST EQUATION

For the simple electrochemical cell consisting of zinc and copper half-cells, we can calculate the free energy change ΔG associated with the redox reaction according to eq. 5.9:

$$\Delta G = \Delta G^0 + RT \ln \frac{\left[Zn^{2+}\right][Cu]}{[Zn]\left[Cu^{2+}\right]} \qquad \text{eq. 5.9}$$

By combining eq. 5.9 with the relationships of ΔG and E or ΔG^0 and E^0 (eq. 5.6 and eq. 5.7), we obtain for the electromotive force E of the cell

$$E = E^0 - \frac{RT}{nF} \ln \frac{\left[Zn^{2+}\right][Cu]}{[Zn]\left[Cu^{2+}\right]} \qquad \text{eq. 5.10}$$

At this point it is again important to remember that the concentrations in eq. 5.9 and eq. 5.10 are concentrations relative to the standard state. The solid electrodes are in their standard concentration, the concentration of the pure metals. Therefore, the terms for the electrodes in the argument of the logarithm have a value of unity, and can be neglected. For ions in solution, the standard concentration is 1 M. ΔG and the electromotive force E therefore depend only on the ratio of the Zn^{2+} and Cu^{2+} concentrations in the half-cells: the standard concentration cancels. Eq. 5.9 thus leads to the same expression for the electromotive force that we have derived from considerations of the electrochemical potential (eq. 5.8).

Scheme 5.12 is a general scheme for the reactions in any electrochemical cell that contains a half-cell in which species 1 is oxidized and a half-cell where species 2 is reduced

$$\begin{array}{r} \mathrm{red_1} \rightleftharpoons \mathrm{ox_1} + n\,\mathrm{e^-} \\ \mathrm{ox_2} + n\,\mathrm{e^-} \rightleftharpoons \mathrm{red_2} \\ \hline \mathrm{red_1} + \mathrm{ox_2} \rightleftharpoons \mathrm{ox_1} + \mathrm{red_2} \end{array} \qquad \text{scheme 5.12}$$

In analogy to eq. 5.9, we can express ΔG for the cell reaction as

$$\Delta G = \Delta G^0 + RT \ln \frac{[ox_1][red_2]}{[red_1][ox_2]} \qquad \text{eq. 5.11}$$

and obtain the general form of the electromotive force as

$$E = E^0 - \frac{RT}{nF} \ln \frac{[ox_1][red_2]}{[red_1][ox_2]} \qquad \text{eq. 5.12}$$

This equation is equivalent to

$$E = E^0 + \frac{RT}{nF} \ln \frac{[red_1][ox_2]}{[ox_1][red_2]} \qquad \text{eq. 5.13}$$

and is called the *Nernst equation*.

For half-cell reactions (Scheme 5.13)

$$\mathrm{ox} + n\,\mathrm{e^-} \rightleftharpoons \mathrm{red} \qquad \text{scheme 5.13}$$

we can write by analogy

$$\Delta G = \Delta G^0 + RT \ln \frac{[red]}{[ox]} \qquad \text{eq. 5.14}$$

and

$$E = E^0 - \frac{RT}{nF} \ln \frac{[red]}{[ox]} \qquad \text{eq. 5.15}$$

or

$$E = E^0 + \frac{RT}{nF} \ln \frac{[red]}{[ox]} \qquad \text{eq. 5.16}$$

We can calculate the potential difference E for the $Zn/Zn^{2+}//Cu^{2+}/Cu$ electrochemical cell (Figure 5.1) directly from the individual electrode potentials of the half-cells. From the electrode potential of the half-cell in which the reduction reaction occurs, $E_{reduction}$, and the electrode potential from the half-cell in which the oxidation reaction occurs, $E_{oxidation}$, we obtain E as

$$E = E_{reduction} - E_{oxidation} \qquad \text{eq. 5.17}$$

It is important to note that the electrode potentials $E_{reduction}$ and $E_{oxidation}$ in eq. 5.17 refer to the tabulated values for the reduction reactions. For the $Zn/Zn^{2+}//Cu^{2+}/Cu$ electrochemical cell, the standard electrode potential E^0 for the Zn half-cell is –0.76 V, for the Cu half-cell E^0 is 0.34 V (Table 5.1). According to eq. 5.17, the potential difference E^0 of the electrochemical cell under standard conditions is 0.34 V–(–0.76 V) = 1.1 V. This positive value corresponds to a negative value for ΔG^0 (eq. 5.7), consistent with a spontaneous reaction.

Eq. 5.17 is also valid when the number of electrons that are taken up or released in the oxidation and reduction reactions differs. In these cases, to write the overall reaction scheme and to calculate ΔG for these types of reactions, we have to multiply the stoichiometry coefficients of the reactions by the appropriate integer to arrive at the same number of electrons released in the oxidation and taken up in the reduction. However, the electrode potentials are proportional to the ratio of ΔG and n ($E = -\Delta G/nF$, see eq. 5.6), and remain unchanged. The potential difference E can thus again be calculated directly from the tabulated values according to eq. 5.17.

We can also calculate the $\Delta G^{(0)}$ values for the complete cell reaction from the $\Delta G^{(0)}$ values of the individual reactions in the half-cells. Here we need to take into account what reaction is taking place in the half-cells. In our $Zn/Zn^{2+}//Cu^{2+}/Cu$ example, Zn is oxidized to Zn^{2+} in one half-cell. The standard electrode potential is –0.76 V for the reduction, corresponding to a positive ΔG^0 and a non-spontaneous reaction under standard conditions. For the reverse reaction, the oxidation of Zn, the electrode potential has the opposite sign, and E^0 = +0.76 V, which we can convert into a ΔG^0 of –176.7 kJ mol^{-1} according to eq. 5.7. In the Cu half-cell, the reduction reaction is taking place, and the electrode potential corresponds to the tabulated value of E^0 = 0.34 V, which can be converted to ΔG^0 = –65.6 kJ mol^{-1}. The ΔG^0 of the overall cell reaction is the sum of the ΔG^0 values for the half-cell reactions, –212.3 kJ mol^{-1}. This value corresponds to an E^0 of 1.1 V for the electrochemical cell, the same value we have calculated from the electrode potentials of the half-cells according to eq. 5.17. However, the electrode potential is not a state function, and electrode potentials are not additive! Whenever we want to calculate the standard electrode potential for an overall reaction from the standard electrode potentials of the individual reactions, we have to take a detour over the ΔG values that are additive, and then convert ΔG for the overall reaction into E (Box 5.1).

BOX 5.1: NON-ADDITIVITY OF *E*.

We consider a reduction reaction that occurs in two steps, the reduction of nitrate ions to nitrite, and the subsequent reduction of nitrite to ammonium (Scheme 5.14). The standard electrode potentials for the individual reactions are $E^0_1 = 0.42$ V and $E^0_2 = 0.48$ V.

$$\begin{array}{lll} NO_3^- + 2\,H^+ + 2\,e^- \longrightarrow NO_2^- + H_2O & & E^0_1 = 0.42\ V \\ NO_2^- + 8\,H^+ + 6\,e^- \longrightarrow NH_4^+ + 2\,H_2O & & E^0_2 = 0.48\ V \\ \hline NO_3^- + 10\,H^+ + 8\,e^- \longrightarrow NH_4^+ + 3\,H_2O & & E^0_{tot} = ? \end{array}$$

scheme 5.14

To calculate the overall standard electrode potential, we convert E^0_1 and E^0_2 into ΔG^0 values using eq. 5.7, and obtain $\Delta G^0_1 = -81.06$ kJ mol^{-1} (with $n = 2$) and $\Delta G^0_2 = -277.92$ kJ mol^{-1} (with $n = 6$). The change in free energy for the overall reaction is the sum of these values, $\Delta G^0_{tot} = -359.0$ kJ mol^{-1}. From ΔG^0_{tot}, we can calculate $E^0_{tot} = 0.465$ V.

From eq. 5.7 and the additivity of ΔG, we can derive a general expression for the electrode potential E_{tot} of the overall reaction as a function of the electrode potentials E_1 and E_2 for the individual reactions:

$$E_{tot} = \frac{n_1 E_1 + n_2 E_2}{n_{tot}} \qquad \text{eq. 5.18}$$

5.5 MEASURING pH VALUES

Using electrochemical cells, pH values of solutions can be measured. As a reference cell, either the hydrogen electrode (Scheme 5.7) or the glass electrode is used. The electrode potential of these electrodes depends directly on the concentration of H^+ ions, and thus of the pH of the test solution. According to the Nernst equation (eq. 5.16), the electrode potential of the H^+/H_2 half-cell is described by:

$$E = E^0_{H^+/H_2} + \frac{RT}{2F} \ln \frac{\left[H^+\right]^2}{p_{H_2}} \qquad \text{eq. 5.19}$$

The standard potential for the H^+/H_2 half-cell is zero. The pressure p_{H_2} is the standard pressure, and can be neglected. (Remember that p_{H_2} is short for the actual pressure relative to the standard pressure p^0). The term $\ln[H^+]^2$ equals $2{\cdot}\ln[H^+]$, and the factor 2 in numerator and denominator of the pre-factor cancels, leaving

$$E = \frac{RT}{F} \ln\left(\left[H^+\right]\right) \qquad \text{eq. 5.20}$$

Converting the natural into the decadic logarithm by multiplying with ln10 = 2.303 (see Section 31.5) gives

$$E = \frac{2.303RT}{F} \log\left(\left[H^+\right]\right) \qquad \text{eq. 5.21}$$

With the definition of the pH as the negative decadic logarithm of the proton concentration, we finally arrive at

$$E = -\frac{2.303RT}{F} pH \qquad \text{eq. 5.22}$$

Eq. 5.22 describes the pH dependence of the electrode potential of the H^+/H_2 half-cell. To determine the pH value of a test solution, the hydrogen electrode is used in combination with a reference electrode with pH-independent electrode potential, and the pH can be calculated from the electromotive force.

Hydrogen electrodes are not very convenient to work with, and glass electrodes are more commonly used for pH determination. Here, the potential at the measurement electrode is generated by the difference in proton concentrations between a reference solution ($[H^+]_{ref}$, typically pH 7) and the test solution of unknown pH ($[H^+]_{test}$), separated by a glass membrane. The potential E at the glass membrane is

$$E = \frac{RT}{F} \ln \left(\frac{\left[H^+ \right]_{test}}{\left[H^+ \right]_{ref}} \right) \qquad \text{eq. 5.23}$$

or

$$E = \frac{2.303RT}{F} \left(pH_{ref} - pH_{test} \right) = \frac{2.303RT}{F} \Delta pH \qquad \text{eq. 5.24}$$

The potential is measured relative to a silver or a calomel electrode (Schemes 5.10 and 5.11) that serves as an internal reference electrode.

5.6 REDOX REACTIONS IN BIOLOGY

5.6.1 Cysteines and disulfide bonds

The side chain of the amino acid cysteine contains a thiol group. Two cysteines in close proximity can be oxidized to a disulfide, termed cystine. The biochemical standard electrode potential for the reduction of cystine to cysteine is $E^{0\prime} = -0.34$ V (Table 5.2). In the cytosol, reducing conditions are maintained, and intracellular proteins therefore do not contain disulfide bonds. In contrast, secreted proteins are frequently stabilized by disulfide bonds (Section 18.1.4.5).

Glutathione is a thiol-containing peptide in the cell of the sequence glutamate-cysteine-glycine. The peptide bond between glutamate and cysteine is formed between the γ-carboxyl group and the α-amino-group of the cysteine; the peptide bond between cysteine and glycine is a regular peptide bond. Glutathione exists in a monomeric, reduced form (GSH) and a dimeric, oxidized form (GSSG). GSSG is reduced to GSH by the enzyme GSH reductase, which uses NADPH as electron donor.

The GSH/GSSG system serves as a major cellular redox buffer. The biochemical electrode potential of the GSH/GSSG system depends on the concentrations of reduced and oxidized forms of GSH according to the Nernst equation (eq. 5.15)

$$E' = E^{0'} - \frac{RT}{nF} \ln \frac{\left[GSH \right]^2}{\left[GSSG \right]} \qquad \text{eq. 5.25}$$

with $n = 2$, and a biochemical standard electrode potential of $E^{0\prime} = -0.34$ V (Table 5.2). The ratio of $[GSH]/[GSSG]$ in the cytoplasm exceeds 100 under normal physiological conditions. Under oxidative stress conditions, it is reduced to < 10. In the endoplasmic reticulum (ER), the $[GSH]/[GSSG]$ ratio is in the range of 1–3, making the ER a much more oxidative environment than the cytoplasm.

Glutathione is an antioxidant that protects cells against reactive oxygen species. It also forms mixed disulfide bonds with cysteines, and protects proteins from oxidation. Furthermore, it reduces disulfide bonds in proteins. The reduction of protein disulfide bonds is also catalyzed by the small proteins glutaredoxin and thioredoxin. Both proteins contain two catalytic cysteines. One of these cysteines reacts with the disulfide bond to be reduced, forming a mixed disulfide. The second cysteine then reacts with the mixed disulfide, leading to the formation of glutaredoxin in the oxidized state, with a disulfide bond, and a reduced target protein, with two cysteines. Subsequently, oxidized glutaredoxin is reduced by GSH in a non-enzymatic reaction, followed by reduction of GSSG by GSH reductase. Oxidized thioredoxin is reduced by thioredoxin reductase with NADPH as reducing agent.

5.6.2 The respiratory chain

The oxidation of metabolites provides reduction equivalents in the form of $NADH/H^+$ and $FADH_2$. These reduction equivalents are oxidized in the respiratory chain in the inner mitochondrial

membrane (Figure 5.2). Molecular oxygen is the final electron acceptor, and is reduced to H_2O. The electrons from NADH/H^+ and $FADH_2$ are not directly transferred to oxygen, however, but pass a chain of electron acceptor/donor pairs. In a first reaction, the NADH/ubiquinone oxidoreductase, a transmembrane protein, catalyzes the oxidation of NADH, and the electron transfer to ubiquinone *via* flavin mononucleotide and an iron-sulfur cluster. In the second reaction, the ubiquinone/cytochrome c oxidoreductase catalyzes the electron transfer from ubiquinone to cytochrome c *via* cytochrome b, an iron-sulfur cluster, and cytochrome c_1. Finally, the cytochrome c/O_2 oxidoreductase catalyzes the electron transfer to molecular oxygen *via* cytochromes a and a_3. NADH,H^+/NAD^+ has the most negative electrode potential (–0.32 V), the potential of O_2/H_2O is the most positive in the chain (0.81 V; Table 5.2). Each oxidation step in the respiratory chain goes from negative to more positive electrode potential, and is exergonic. All three oxidoreductases couple the energy from electron transport to the endergonic transport of protons into the intermembrane space, establishing a proton gradient (Section 4.2; Figure 4.4). The electrode potential of $FADH_2$/FAD (–0.22 V; Table 5.2) is less negative than the NADH potential, and oxidation of $FADH_2$ provides less energy. Electrons from $FADH_2$ are channeled into the respiratory chain in the second reaction catalyzed by the ubiquinone/cytochrome c reductase. The oxidation of $FADH_2$ is thus only coupled to two proton pumps, and it contributes less to the establishment of the proton-motive force.

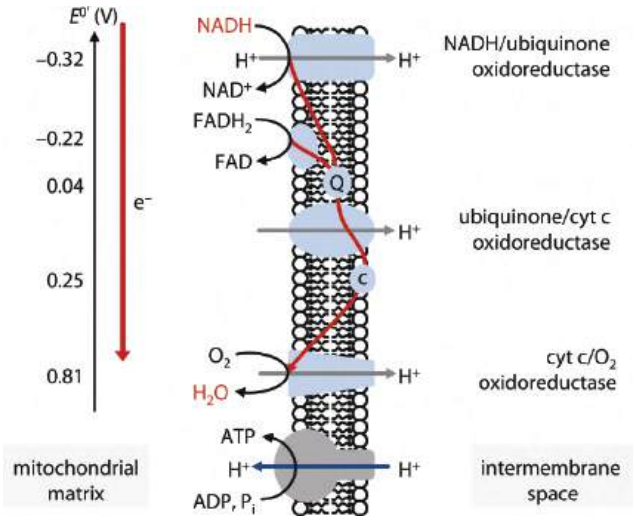


FIGURE 5.2: Electron transport in the respiratory chain. Electrons from oxidation of NADH to NAD^+ are transferred through a chain of electron acceptors and donors to oxygen that is reduced to H_2O. Electron transport is mediated by the transmembrane complexes NADH/ubiquinone reductase, ubiquinone/cytochrome c reductase and cytochrome c/O_2 reductase (light blue). The electron flux is from negative to positive electrochemical potential (red arrow) through a series of exergonic reactions. Electron transfer is coupled to the endergonic transport of protons from the mitochondrial stroma to the intermembrane space, mediated by the transmembrane complexes. The proton transport generates a concentration gradient. Influx of protons through the F_1-F_o ATP synthase (dark gray) leads to ATP synthesis. Electrons from $FADH_2$ are channeled into the transport chain after NADH/ubiquinone reductase, hence these electrons contribute less to the proton gradient and ATP synthesis. Flavin mononucleotide, cytochromes, and iron-sulfur clusters that serve as electron acceptors and donors within the proton-pumping complexes are not depicted. The electrode potential $E^{0\prime}$ is not to scale. Q: ubiquinone, cyt c: cytochrome c.

5.6.3 The Light Reaction in Photosynthesis

During photosynthesis, absorption of light by chlorophyll molecules in the reaction centers of photosystems I and II (PS I, PS II) induces electronic transitions (see Section 21.1), converting these chlorophylls into a highly reducing form with a negative electrode potential. The reaction center of photosystem I, P_{700}, transfers its electrons in an electron transport chain *via* ferredoxin to the

terminal electron acceptor $NADP^+$, whose reduction to NADPH is catalyzed by the ferredoxin/$NADP^+$ oxidoreductase. The oxidized P_{700} has to be converted back into the reduced form, which is achieved by an electron transport chain from the oxidized reaction center of photosystem II, P_{680}. P_{680} transfers its electrons to P_{700} *via* plastoquinone, the cytochrome bf complex and plastocyanin. P_{680} is reduced to its original state by electron transfer from water. The exergonic electron transport from H_2O to $NADP^+$ is coupled to the endergonic establishment of a proton-motive force that is used for ATP synthesis (Section 4.2; Figure 4.4).

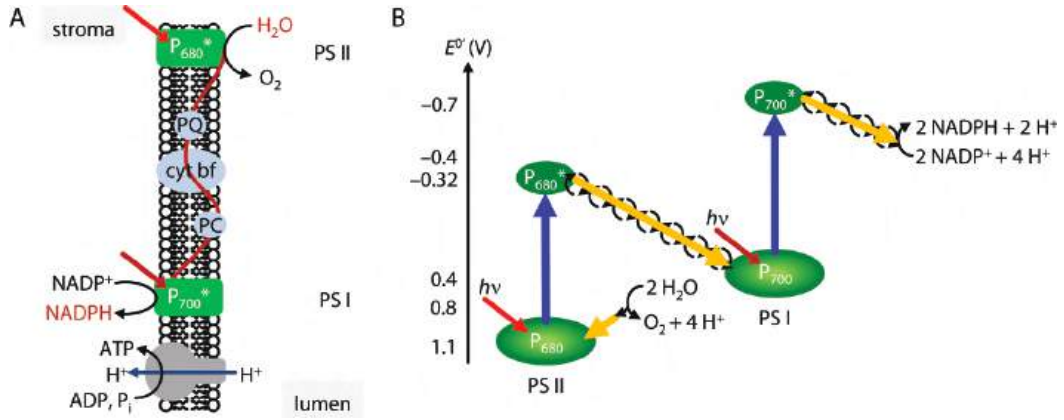


FIGURE 5.3: Electron transport in the light reaction of photosynthesis. A: Arrangement of the photosystems I and II and the electron carriers in the chloroplast membrane. Electron transport from H_2O to $NADP^+$ is coupled to proton transport, and the proton-motive force drives ATP synthesis. PQ: plastoquinone, PC: plastocyanin. B: Electron transport chain. Excitation of P_{700} in PS I by light changes the redox potential, and P_{700}* can deliver electrons to $NADP^+$ through a transport chain of electron acceptors and donors. PS II (P_{680}) is excited by light and transports electrons to PS I through a transport chain. PS II is reduced to the starting state by electrons from H_2O. The exergonic electron flux from negative to positive potential (orange arrows) is coupled to the transport of protons into the lumen. The generated proton-motive force is used for ATP synthesis.

To maintain metabolic processes, the NAD^+/NADH ratio has to be balanced. A large portion of hangover symptoms after excessive alcohol consumption has been ascribed to the imbalance of the NAD^+/NADH ratio because of the NAD^+ consumption in the oxidation of alcohol.

5.7 THE ELECTROCHEMICAL POTENTIAL AND MEMBRANE POTENTIALS

Biological membranes are impermeable for ions, and ion transport across the membrane is regulated (Section 4.3). We have discussed before that concentration gradients are useful to store energy (Section 4.2). To calculate the energy stored in an ion gradient, we have to take into account the energy stored in the concentration difference, plus the energy associated with the charge separation. We have defined the electrochemical potential of ions in the presence of electric fields by correcting their chemical potential μ for the electrochemical work (see eq. 5.1 at the very beginning of this Chapter). In the same way that the chemical potential is the driving force for a chemical reaction, the electrochemical potential is the driving force for an electrochemical reaction. If the electrochemical potential on the inside of a membrane is larger than the electrochemical potential on the outside, then the efflux of ions is the thermodynamically favored reaction:

$$\tilde{\mu}_{j,in} > \tilde{\mu}_{j,out} \qquad \text{eq. 5.26}$$

Thus, ions would spontaneously pass the membrane from inside to outside if the membrane were ion-permeable. Conversely, if

$$\tilde{\mu}_{j,in} < \tilde{\mu}_{j,out} \qquad \text{eq. 5.27}$$

then the spontaneous flow would be from the outside to the inside, and influx of ions is thermodynamically favored. Equilibrium is reached when the electrochemical potential on both sides of the membrane is equal:

$$\tilde{\mu}_{j,in} = \tilde{\mu}_{j,out} \qquad \text{eq. 5.28}$$

The impermeability of biological membranes for ions allows the establishment and maintenance of ion gradients. By regulating opening and closing of ion channels in the membrane, the membrane potential can be modified. The transmission of neuronal signals is achieved by changes in the membrane potential of neurons. We will first derive the equation describing a membrane potential considering only sodium ions. In equilibrium, their electrochemical potential on both sides of the membrane must be equal:

$$\tilde{\mu}_{Na^+,in} = \tilde{\mu}_{Na^+,out} \qquad \text{eq. 5.29}$$

From the definition of the electrochemical potential (eq. 5.1), we obtain

$$\mu^0_{Na^+} + RT \ln c_{Na^+,in} + z_{Na^+} F \phi_{Na^+,in} = \mu^0_{Na^+} + RT \ln c_{Na^+,out} + z_{Na^+} F \phi_{Na^+,out} \qquad \text{eq. 5.30}$$

The charge of the sodium ions is 1, and z_{Na^+} can be neglected. We can rearrange eq. 5.30, and can calculate the potential difference $\Delta\phi$ as

$$\Delta\phi = \phi_{in} - \phi_{out} = \frac{RT}{F} \ln \frac{c_{Na^+,out}}{c_{Na^+,in}} \qquad \text{eq. 5.31}$$

If more ions contribute to a membrane potential, eq. 5.31 becomes more complex because all relevant ions have to be taken into account. For the membrane potential of a neuron, Na^+ and K^+ are the most relevant ions. In the resting state of a neuron, the concentration of Na^+ ions is 150 mM on the outside and 15 mM on the inside. The K^+ concentration is 5 mM on the outside, and 140 mM on the inside. These concentrations are maintained by the activity of the Na^+/K^+ transporter, an antiporter that transports 3 Na^+ ions out of and 2 K^+ ions into the cell at the expense of ATP hydrolysis. From eq. 5.31, we can calculate the equilibrium potential at 37°C from the Na^+ gradient as +62 mV (positively charged on the outside, negative on the inside), for the K^+ gradient it is –90 mV (negatively charged on the outside, positive on the inside). The membrane potential of a neuron in the resting state is typically –60 to –80 mV (Figure 5.4), and thus much closer to the K^+ potential. This is caused by potassium channels that are open and permit flux of K^+ from the inside to the outside, along the concentration gradient. In contrast, Na^+ channels are closed in the resting state. The overall membrane potential is therefore predominantly determined by the K^+ equilibrium potential.

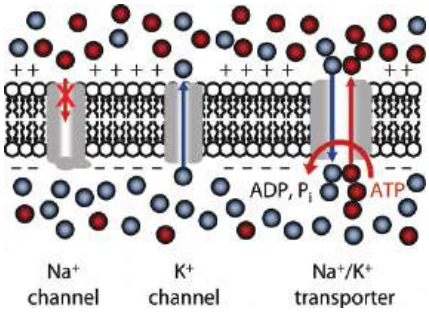


FIGURE 5.4: Membrane potential of a neuron in the resting state. The Na^+/K^+ transporter maintains the concentration gradient of Na^+ and K^+ ions across the neuronal membrane by actively transporting Na^+ ions out of and K^+ ions into the cell against the concentration gradient. Some K^+ channels are open and permit efflux of K^+. The resulting membrane potential is closer to the K^+ than to the Na^+ equilibrium potential.

For more than two ions, the membrane potential can be calculated from the concentrations of all cations M_i^+ and anions X_i^- on the outside, multiplied with the (relative) permeability P_i, in the numerator of the logarithmic term, and the concentrations of all cations M_i^+ and anions X_i^- on the inside, multiplied with the permeability P_i in the denominator. The membrane potential is

$$\Delta\phi = \frac{RT}{F} \ln \frac{\sum P_i \left[M_i^+\right]_{\text{outside}} + \sum P_i \left[X_i^-\right]_{\text{outside}}}{\sum P_i \left[M_i^+\right]_{\text{inside}} + \sum P_i \left[X_i^-\right]_{\text{inside}}} \qquad \text{eq. 5.32}$$

Eq. 5.32 implies that the membrane potential changes if the relative permeabilities of one (or more) ions change. This is indeed the mechanism of neuron activation and the generation and transmission of electric signals in the neural system. The permeability is altered by the opening and closing of gated channels in the membrane in response to a certain signal. Opening of K^+ channels allows increased efflux of K^+ from the cell, and the membrane potential becomes more negative. This effect is called *hyperpolarization*. On the other hand, opening of Na^+ channels enables increased influx of Na^+, and reduces the negative charge on the inside (*depolarization*). Many of the gated channels involved in these processes are voltage-dependent, meaning they open and close in response to changes in membrane potential. Voltage-dependent Na^+ channels open in response to a slight depolarization of the membrane. Influx of Na^+ ions leads to further depolarization and opening of more Na^+ channels in a positive feedback loop, until a threshold value of the potential is reached. Quickly, all voltage-gated Na^+ channels will be open, and the membrane potential approaches the Na^+ equilibrium potential: an *action potential* has been generated (Figure 5.5).

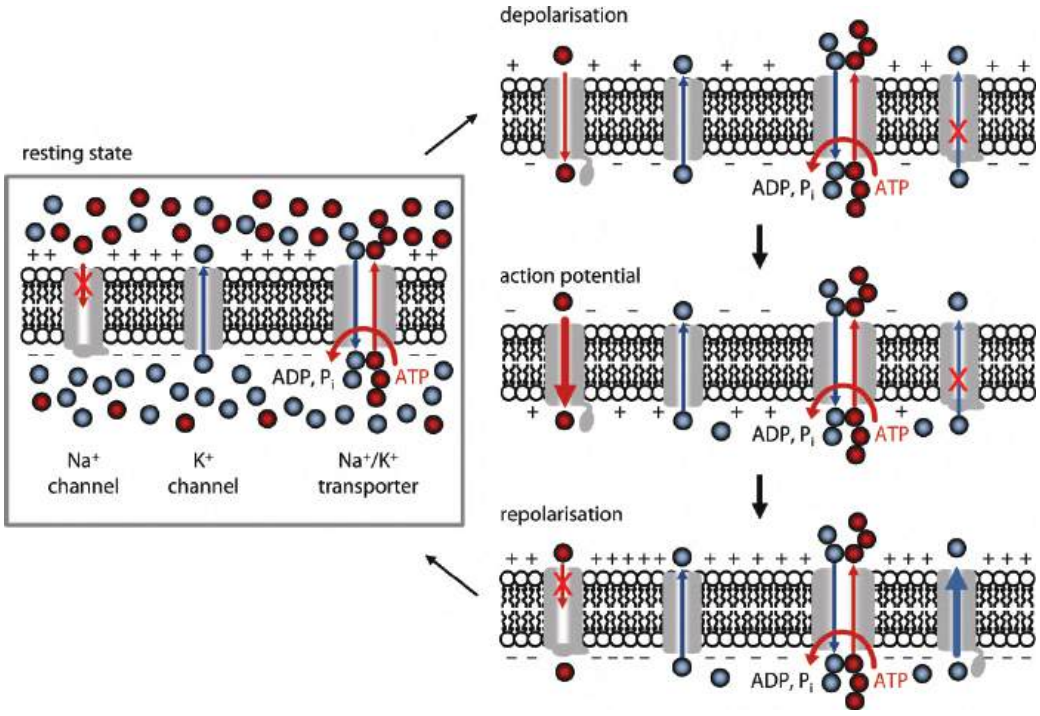


FIGURE 5.5: Generation and decline of an action potential. When a signal leads to a slight depolarization, voltage-gated Na^+ channels open, and influx of sodium ions leads to a further depolarization of the cell. Before the Na^+ equilibrium potential has been reached, the Na^+ channels become inactivated, and voltage-gated K^+ channels open. The efflux of K^+ leads to a repolarization of the membrane. The membrane permeability for K^+ is higher than in the resting state, leading to a transient hyperpolarization, before the voltage-dependent K^+ channels close and the resting potential is re-established.

This action potential is localized because the original depolarization was localized to a defined region of the neuron. The Na^+ channels only open for a short period of time, however, and are inactivated before the Na^+ equilibrium potential has been reached. At the same time, voltage-gated K^+ ion channels open, and K^+ ion flux out of the cell shifts the membrane potential back toward the K^+ equilibrium potential: the membrane becomes repolarized. With the voltage-gated K^+ channels open, the permeability for K^+ ions is larger than in the resting state. The membrane potential will therefore briefly overshoot and temporarily become more negative, before the equilibrium membrane potential of the resting state is reached. As soon as the Na^+ channels can be activated again, a new action potential can be generated. The transient inactivation of Na^+ channels ensures that the action potential moves along the axon in one direction, away from the area where it was generated (Figure 5.5).

5.8 ELECTROPHYSIOLOGY: PATCH-CLAMP METHODS TO MEASURE ION FLUX THROUGH ION CHANNELS

In Section 5.7, we have seen that controlled opening and closing of ion channels alters the ion permeability and thus the electrical conductivity of biological membranes. The current through these ion channels can be measured by the *patch-clamp* technique (Figure 5.6). In a patch-clamp experiment, a glass pipette with a micrometer opening is manually brought into contact with the membrane of a cell under visual control in a microscope. Suction is applied, which leads to a tight sealing of the glass walls at the opening with the membrane, while the cell remains intact (*cell-attached mode*). The glass-membrane contact has a high resistance and low conductivity, and prevents leakage of ions between the inside and outside of the micropipette. The part of the membrane that closes off the opening of the micropipette is called the *patch*. A flux of ions between inside and outside is possible if one or more ion channels are located within this patch, and can be measured as a current. This approach allows one to measure picoampere currents through single ion channels in their natural environment. For voltage-gated ion channels, different voltages can be applied, and the current-voltage response curve of a single channel can be measured. By patch-clamping a ligand-gated ion channel in the cell-attached mode, the response of the channel to different ligands can be tested. By measuring the current through the channel at different ligand concentrations, a dose-response curve for a particular ligand may be determined. In the cell-attached mode, the cytoplasmic solution cannot be manipulated because the cell remains intact. Therefore, the ligand has to be supplied with the solution inside the micropipette, and so cannot be changed after patching. Hence, the dose-response curve has to be measured point by point by clamping a new patch for each ligand concentration. By gently pulling on the membrane-sealed micropipette, the patch can be separated from the rest of the membrane. The cytosolic side of the membrane now faces to the outside of the micropipette, hence the name *inside-out configuration*. The solution outside the micropipette can easily be changed, and the inside-out configuration is therefore suitable to study the effect of intracellular, cytoplasmic ligands on ligand-gated ion channels. To study the effect of extracellular ligands such as neurotransmitters and drugs on ligand-gated channels, the *outside-out configuration* is used. To achieve this, more suction is applied in the cell-attached mode, which leads to opening of the membrane patch. The seal between pipette and membrane remains intact, and the solutions inside and outside the cell remain separated, but the solution in the micro-pipette is now in contact with the cytoplasm. In this *whole-cell mode*, the current through all channels in the entire cell membrane is measured. By pulling on the micropipette, the cell membrane ruptures, and membrane fractions sealed to the glass can form patches where the extracellular face of the membrane faces outside, hence the name outside-out configuration. In this mode, the response of single ion channels to extracellular ligands and their dose-response curves can be measured.

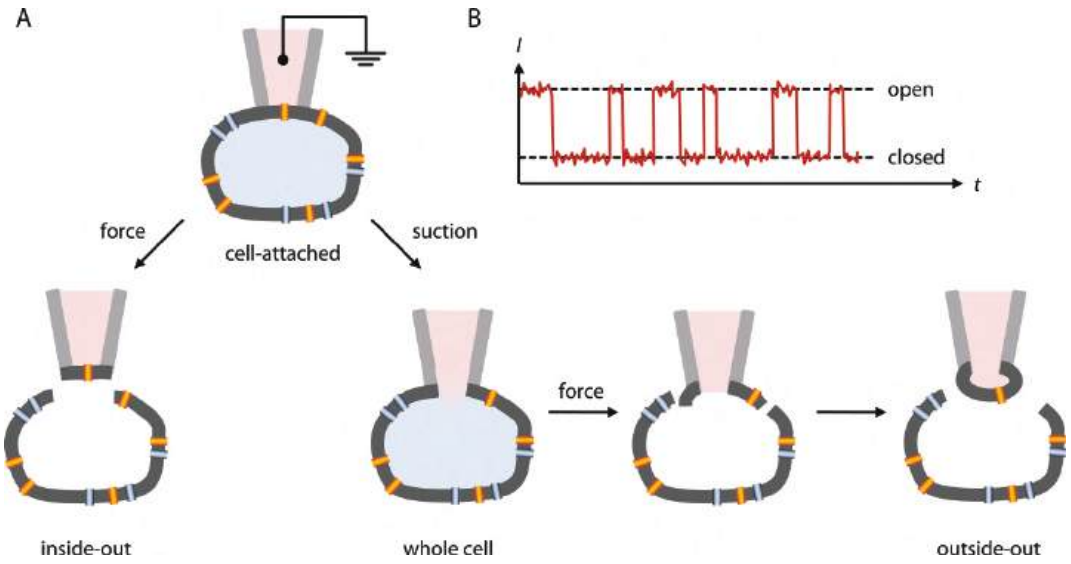


FIGURE 5.6: Measurement of ion flux through ion channels by patch-clamp techniques. A: Different patch-clamp configurations. A micropipette (gray/light red) is moved onto the surface of a cell and tightly attaches to the membrane upon suction. By applying force, the patch attached to the pipette can be extracted from the rest of the membrane. The intracellular (cytosolic) side of the membrane faces toward the outside of the pipette (inside-out). By increasing the suction instead of pulling, the patch ruptures, and the micropipette is sealed to the cell (whole cell). By force, the membrane can be ruptured. Membrane fragments sealed to the micropipette form a new patch. The extracellular side of the membrane faces the outside of the micropipette (outside-out). B: Schematic measurement: current *I* (in pA) as a function of time (in s). When the ion channel is open, the current is high. When the channel closes, no current is measured.

The patch-clamp method, developed by Neher and Sakmann in 1976, has more recently been combined with fluorescence microscopy to directly monitor the binding of fluorescently labeled ligands, conformational changes or subunit stoichiometry in parallel to measuring the current.

QUESTIONS

5.1 A pH meter provides a readout of 0 V for a solution of pH 7 at 25°C. What range of voltages is measured for pH 1–14?

5.2 An electrochemical cell consists of Zn in a $ZnSO_4$ solution and Cu in a $CuSO_4$ solution. The $CuSO_4$ solution has a concentration of 100 mM. What would be the minimal concentration of a $ZnSO_4$ solution to drive the reaction in the Cu^{2+}/Cu half-cell toward Cu^{2+}? E^0 (Zn^{2+}/Zn) = −0.76 V, E^0 (Cu^{2+}/Cu) = 0.34 V.

5.3 The equilibrium constant $K = Cu^{2+}/(Cu^{+})^2$ in aqueous solution is 10^6 M^{-1}. The standard potential of a Cu/Cu^{2+} electrode is E^0 = −0.34 V. What is the standard potential for the Cu^{+}/Cu electrode and for the Cu^{2+}/Cu^{+} electrode?

5.4 The standard potential for the reaction $NAD^{+} + H^{+} + 2\ e^{-} \rightarrow NADH$ is −0.099 V at 20°C and−0.105 V at 30°C. Calculate the change in standard free energy, entropy, and enthalpy for 25°C, assuming a linear dependence of E^0 on temperature.

5.5 What is $E^{0'}$ for an NAD^{+}/H^{+}/NADH-Pt half-cell at pH 7? E^0 = −0.11 V.

5.6 A solution contains pyruvate, lactate, NAD^{+}, and NADH. The concentration ratios are pyruvate/lactate = 1, NAD^{+}/NADH = 1. What reactions do occur, what is E^0 and what is $\Delta G^{0'}$? How do these values change for a 500-fold excess of lactate over pyruvate and of NAD^{+} over NADH? At what ratio of lactate/pyruvate = NAD^{+}/NADH are the reactions in equilibrium? $E^{0'}$(NAD+/NADH) = −0.32 V; $E^{0'}$(pyruvate/lactate) = −0.19 V.

5.7 The biochemical electrode potential of the GSH/GSSG system depends on the concentrations of reduced and oxidized forms of GSH according to the Nernst equation.

$$E' = E^{0'} - \frac{RT}{nF} \ln \frac{[GSH]^2}{[GSSG]}$$

with n = 2, and a biochemical standard electrode potential of $E^{0'}$ = −0.34 V. The total cellular concentration of GSH is approx. 5 mM. The ratio of [*GSH*]/[*GSSG*] in the cytoplasm exceeds 100 under normal physiological conditions. Under oxidative stress conditions, it is reduced to < 10. In the endoplasmic reticulum (ER), the [*GSH*]/[*GSSG*] ratio is in the range of 1–3. Calculate E' for all three conditions and rank them according to their reducing character. T = 37°C, F = 96500 C.

5.8 Calculate the membrane potential of a cell for a concentration of K^{+} ions that is 20-fold higher inside than outside. What membrane potential results from a Na^{+} concentration that is 10-fold higher outside than inside? T = 37°C.

5.9 *E. coli* maintains a ΔpH of 1 (inside − outside) and a $\Delta\phi$ of −120 mV across its cell membrane. What is the energy derived from this concentration gradient, and what is the maximum concentration of a metabolite M inside *versus* outside that can be achieved by its symport into the cell with H^{+}? What is the maximum concentration if M is negatively charged (M^{-})?

5.10 Mitochondria maintain a ΔpH of 1.4 (inside − outside) and a $\Delta\phi$ of −0.14 V (inside − outside). How many protons need to pass the membrane for synthesis of one ATP molecule? $\Delta G^{0'}$ (ATP synthesis) = 31 kJ mol^{-1}.

REFERENCES

Gilbert, H. F. (1984) Redox control of enzyme activities by thiol/disulfide exchange. *Methods Enzymol* **107**: 330–351.

Gilbert, H. F. (1995) Thiol/disulfide exchange equilibria and disulfide bond stability. *Methods Enzymol* **251**: 8–28.
· describe the ratio of GSH and GSSG under normal and oxidative stress conditions

Hwang, C., Sinskey, A. J. and Lodish, H. F. (1992) Oxidized redox state of glutathione in the endoplasmic reticulum. *Science* **257**: 1496–1502.
· reports the ratio of GSH and GSSG in the endoplasmic reticulum

Kusch, J. and Zifarelli, G. (2014) Patch-clamp fluorometry: Electrophysiology meets fluorescence. *Biophys J* **106**(6): 1250–1257.
· review about the combination of patch-clamp methods with fluorimetry to study ion channel conformational changes and ligand binding

Neher, E. and Sakmann, B. (1976) Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature* **260**(5554): 799–802.
· original paper on the patch-clamp method

Talwar, S. and Lynch, J. W. (2015) Investigating ion channel conformational changes using voltage clamp fluorometry. *Neuropharmacology* **98**: 3–12.
· review about the combination of voltage-clamp methods with fluorimetry to study conformational dynamics of ion channels



PART II
Kinetics

In Part I, we have discussed energy changes associated with changes in the state of a system, transport processes, and chemical reactions. We have so far not considered the velocity with which these reactions occur. In the following Chapters, we will now derive concepts to describe the velocities of increasingly complex reactions, and to understand how enzymes catalyze biochemical reactions.

6

Reaction velocities and rate laws

During a chemical reaction, the concentration of reactants decreases, and the concentration of products increases. The *reaction velocity* v is defined as the change in concentration of the reactants or products with time (Figure 6.1):

$$v = \frac{dc}{dt} = -\frac{d[reactants]}{dt} = \frac{d[products]}{dt} \qquad \text{eq. 6.1}$$

The velocity of a reaction is always positive. For the definition of the reaction velocity in terms of the decrease of reactant concentration we therefore need a negative sign to convert it into a positive value. From eq. 6.1, we can derive the unit of the reaction velocity v as M s^{-1}.

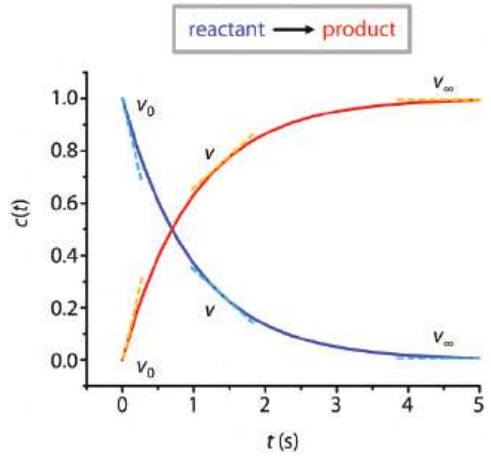


FIGURE 6.1: Concentration changes and reaction velocity. The concentration of the reactant (blue) decreases with reaction time, whereas the concentration of product (red) increases. The reaction velocity is the change in concentration with time. Broken lines indicate the initial velocity v_0, the velocity v at time t, and the velocity v_∞ at the end of the reaction. The reaction velocity decreases from the initial velocity v_0 to the final velocity $v_\infty = 0$ during the reaction.

For a reaction of A and B to C (Scheme 6.1),

$$A + B \longrightarrow C \qquad \text{scheme 6.1}$$

we can express the reaction velocity as the change in concentration of A, B, or C as

$$v = \frac{dc}{dt} = -\frac{d[A]}{dt} = -\frac{d[B]}{dt} = \frac{d[C]}{dt} \qquad \text{eq. 6.2}$$

In the formulation of reaction velocities, the stoichiometry coefficients of the reactants and products need to be taken into account. For a similar reaction of A and two B to C (Scheme 6.2),

$$A + 2\,B \longrightarrow C \qquad \text{scheme 6.2}$$

the reaction velocity is

$$v = \frac{dc}{dt} = -\frac{d[A]}{dt} = -\frac{1}{2}\cdot\frac{d[B]}{dt} = \frac{d[C]}{dt} \qquad \text{eq. 6.3}$$

The factor ½ in eq. 6.3 is necessary because for each A that reacts, two molecules of B are consumed. The concentration of B therefore decreases twice as fast as the concentration of A. The reaction velocity, on the other hand, must be identical independent of which change in concentration we use as a reference.

To monitor the progress of chemical reactions and to determine reaction velocities, we need to follow changes in the concentrations of reactants or products. This can be done discontinuously by taking samples at different time points, with subsequent analysis, or continuously by online monitoring of a suitable spectroscopic signal that reports on disappearance of reactants or appearance of products. While some methods provide absolute information on reactant or product concentrations, others provide a signal that is proportional to the concentration and only measure relative changes. Methods to monitor chemical reactions will be treated in Part IV, particularly in Chapter 28.

Rate laws describe the dependence of reaction velocities on the concentrations of reactants A, B, C,… . Reaction velocities are proportional to the concentrations of the reactants, and rate laws have the general form

$$v = \frac{dc}{dt} = k[A]^a[B]^b[C]^c\ldots \qquad \text{eq. 6.4}$$

The proportionality constant k in eq. 6.4 is the *rate constant* of the reaction. The exponents a, b, c, … are the *order of the reaction* with respect to the reactants A, B, C, … . They are empirical numbers that are determined experimentally. To determine a, the concentration of reactant A is varied, while $[B]$, $[C]$, … are kept constant, and the reaction velocity is measured as a function of $[A]$. The value of b can be determined by varying the concentration of $[B]$ while keeping $[A]$, $[C]$, … constant, and so on. With the empirical exponents, a rate law in the form of eq. 6.4 can then be formulated. The overall order of the reaction is the sum of the orders for the individual reactants, $a+b+c+\ldots$. Reactions whose velocity is independent of the reactant concentration are called *zero-order reactions*. Such a behavior is observed in catalyzed reactions at a high excess of reactants (see Sections 13.1 and 16.1). *A first-order reaction* depends linearly on the concentration, a *second-order reaction* depends on the square of the concentration. We will treat first- and second-order reactions in more detail in Chapter 7.

The reaction order, overall or with respect to a particular reactant, can be a non-integer number. Many reactions proceed through a sequence of reaction steps that define the *reaction mechanism*. The rate laws of such complex, multi-step reactions cannot be derived from the overall reaction

scheme, but depend on the individual steps of the reaction. Rate laws therefore reflect the mechanism of the reaction.

Reactions that occur the way they are written in the reaction scheme and cannot be further subdivided into a sequence of individual reaction steps are called *elementary reactions*. If the reaction in Scheme 6.3 is an elementary reaction, such that *a* molecules of A, *b* molecules of B, *c* molecules of C, … collide to form the reaction products

$$a\mathrm{A} + b\mathrm{B} + c\mathrm{C} + \ldots \xrightarrow{k} \text{products} \qquad \text{scheme 6.3}$$

then we can write the rate law as

$$v = \frac{\mathrm{d}c}{\mathrm{d}t} = k[A]^a[B]^b[C]^c \ldots \qquad \text{eq. 6.5}$$

This rate law is formally identical to the general rate law in eq. 6.4, but now the exponents are the stoichiometry coefficients *a*, *b*, and *c*. A reaction involving one molecule of A depends linearly on [*A*], a reaction involving two molecules of A depends on $[A]^2$, and so on. For an elementary reaction, the reaction order is thus an integer number.

The number of molecules that react in an elementary reaction is defined as the *molecularity* of the reaction. The molecularity is always an integer number, there are no half molecules. We call reactions *unimolecular* if one reactant molecule is involved, *bimolecular* for two reactants, and *trimolecular* for three reactants. A higher molecularity has not been observed, possibly because it is very unlikely that more than three molecules collide with each other in the correct orientation and with sufficient energy for a reaction to occur. For elementary reactions, the molecularity and the overall order of the reaction are equal. However, this is an exception, and in general it is important to distinguish between molecularity and reaction order.

A simple example of a one-step elementary reaction is the bimolecular reaction of hydrogen with iodine to HI (Scheme 6.4):

$$\mathrm{H_2} + \mathrm{I_2} \xrightarrow{k} 2\,\mathrm{HI} \qquad \text{scheme 6.4}$$

With the information that the reaction is an elementary reaction, we can write the rate law as

$$v = \frac{\mathrm{d}c}{\mathrm{d}t} = \frac{1}{2} \cdot \frac{\mathrm{d}[HI]}{\mathrm{d}t} = k[H_2][I_2] \qquad \text{eq. 6.6}$$

The factor ½ takes into account that the reaction of one H_2 and one I_2 produces two molecules of HI, and the concentration of HI therefore increases twice as fast as the concentrations of H_2 and I_2 decrease (see eq. 6.3). Despite the same overall reaction scheme (Scheme 6.5), the situation is different for the corresponding reaction of bromine with hydrogen:

$$\mathrm{H_2} + \mathrm{Br_2} \xrightarrow{k} 2\,\mathrm{HBr} \qquad \text{scheme 6.5}$$

This reaction follows a radical mechanism and occurs in several steps (Box 6.1). Experimental determination of the reaction velocity yielded the empirical rate law

$$v = \frac{\mathrm{d}c}{\mathrm{d}t} = \frac{1}{2} \cdot \frac{\mathrm{d}[HBr]}{\mathrm{d}t} = \frac{k[H_2]\sqrt{[Br_2]}}{1 + k'\dfrac{[HBr]}{[Br_2]}} \qquad \text{eq. 6.7}$$

The drastically different rate laws for these two formally similar reactions exemplify that the rate law of a reaction cannot be deduced from the overall reaction scheme.

BOX 6.1: MECHANISM OF THE REACTION OF BROMINE WITH HYDROGEN.

Bromine and hydrogen react in a multi-step radical reaction. First, Br_2 decomposes into two bromine radicals:

$$Br_2 \xrightarrow{k_1} 2\,Br\cdot \qquad \text{scheme 6.6}$$

The radicals generated in this starting reaction can then react with molecular hydrogen

$$H_2 + Br\cdot \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} HBr + H\cdot \qquad \text{scheme 6.7}$$

producing hydrogen radicals. These hydrogen radicals also react with Br_2 to give HBr

$$Br_2 + H\cdot \xrightarrow{k_3} HBr + Br\cdot \qquad \text{scheme 6.8}$$

leading to regeneration of a bromine radical, and keeping the chain-reaction going. The reaction stops with the recombination of two bromine radicals, which is the reverse reaction of the chain starting reaction (Scheme 6.6).

$$2\,Br\cdot \xrightarrow{k_{-1}} Br_2 \qquad \text{scheme 6.9}$$

From the set of elementary reactions that define the reaction mechanism, the rate law for HBr formation can be derived as:

$$v = \frac{1}{2}\cdot\frac{\mathrm{d}[HBr]}{\mathrm{d}t} = \frac{k_2\sqrt{\frac{k_1}{k_{-1}}}[H_2]\sqrt{[Br_2]}}{1+\frac{k_{-1}}{k_3}\frac{[HBr]}{[Br_2]}} \qquad \text{eq. 6.8}$$

which is in agreement with eq. 6.7 for

$$k = k_2\sqrt{\frac{k_1}{k_{-1}}} \qquad \text{eq. 6.9}$$

and

$$k' = \frac{k_{-1}}{k_3} \qquad \text{eq. 6.10}$$

The rate law in eq. 6.8 can be derived using the steady-state approximation (Section 13.2), its derivation is a question in Chapter 13. The agreement of the empirical rate law (eq. 6.7) with the rate law derived from the mechanism (eq. 6.8) tells us that this reaction mechanism describes the pathway of the reaction.

QUESTIONS

6.1 Express the reaction velocities of the following reactions in terms of each reactant/product:
(1) $S \rightarrow P$
(2) $E + S \rightarrow E + P$
(3) $ATP + H_2O \rightarrow ADP + P_i$
(4) $NAD^+ + H^+ (+ 2\ e^-) \rightarrow NADH$
(5) $H_2O \rightarrow H^+ + OH^-$
(6) $2M \rightarrow D$

6.2 What is the reaction velocity for a reaction of type $aA + bB \rightarrow cC + dD$?

6.3 Write the rate laws for the forward and reverse reactions:
(1) $S \rightleftarrows P$
(2) $E + S \rightleftarrows E + P$
(3) $ATP + H_2O \rightleftarrows ADP + P_i$
(4) $NAD^+ + H^+ (+ 2\ e^-) \rightleftarrows NADH$
(5) $H_2O \rightleftarrows H^+ + OH^-$
(6) $2M \rightleftarrows D$
The forward rate constant is k_f, the rate constant for the reverse reaction is k_r.

6.4 What is the order of the reaction of iodine and hydrogen with respect to the reactants, and what is the overall order of the reaction? What is the molecularity? What are the corresponding orders and molecularities for the individual steps of the reaction of bromine with hydrogen?

6.5 The concentration of a reactant A decays exponentially according to $[A](t) = A_0 e^{-kt}$ with the rate constant k. What is the reaction velocity as a function of time? What is the initial velocity of the reaction?

7

Integrated rate laws for uni- and bimolecular reactions

We have seen in Chapter 6 that rate laws describing the dependence of reaction velocities on the concentrations of reactants can be formulated directly from the reaction scheme of elementary reactions. By integrating the rate law, the time dependence of the concentration of reactants or products is obtained, the *integrated rate law*. In the following, we will derive integrated rate laws for common types of elementary reactions. We start by considering a simple unimolecular reaction where a compound A decays to products (Scheme 7.1):

$$\mathrm{A} \xrightarrow{k} \text{products} \qquad \text{scheme 7.1}$$

An example of such a reaction is radioactive decay (Box 7.2). The reaction velocity v of such a reaction is defined as the change in the concentration of A over time:

$$v = \frac{\mathrm{d}c}{\mathrm{d}t} = -\frac{\mathrm{d}[A]}{\mathrm{d}t} \qquad \text{eq. 7.1}$$

For the unimolecular reaction according to Scheme 7.1, the reaction velocity v at any given point in time is proportional to the actual concentration of A that can react to products:

$$v \propto [A] \qquad \text{eq. 7.2}$$

The reaction velocity depends on the first power of the concentration of A; the reaction is a first-order reaction. We now introduce the proportionality constant k, the rate constant of the reaction:

$$v = k[A] = -\frac{\mathrm{d}[A]}{\mathrm{d}t} \qquad \text{eq. 7.3}$$

The time-dependent change of v as a function of the concentration of A in eq. 7.3 is the rate law for this reaction. We have seen before that the reaction velocity has units M s^{-1}. If we now compare the units in eq. 7.3, we can deduce that the rate constant k for a first-order reaction is measured in s^{-1}.

Eq. 7.3 is a differential equation, and we can separate the variables A and t by rearranging to

$$\frac{\mathrm{d}[A]}{[A]} = -k\,\mathrm{d}t \qquad \text{eq. 7.4}$$

To obtain an expression of the concentration of A as a function of time, we now integrate from A_0, the starting concentration at time $t = 0$, to the concentration $[A](t)$ at time t on the left-hand side, and from $t = 0$ to t on the right-hand side:

$$\int_{A_0}^{[A](t)} \frac{\mathrm{d}[A]}{[A]} = -k\int_0^t \mathrm{d}t \qquad \text{eq. 7.5}$$

Integration gives

$$\ln\left(\frac{[A](t)}{A_0}\right) = -kt \qquad \text{eq. 7.6}$$

or

$$[A](t) = A_0 e^{-kt} \qquad \text{eq. 7.7}$$

The time dependence of the reactant concentration is the integrated rate law. According to eq. 7.7, the concentration of A decays exponentially from its starting value A_0 at $t = 0$ to zero at the end of the reaction ($t \to \infty$) with the rate constant k (Figure 7.1). The larger k, the faster is the reaction, and the faster A is consumed. A_0 is the *amplitude* of the exponential function, and often called the amplitude of the reaction. If A is converted into one product B, and we start the reaction from the initial concentration A_0 without any B ($B_0 = 0$), we can formulate the mass conservation as

$$A_0 = [A](t) + [B](t) \qquad \text{eq. 7.8}$$

which gives

$$[B](t) = A_0 - [A](t) = A_0 - A_0 e^{-kt} = A_0\left(1 - e^{-kt}\right) \qquad \text{eq. 7.9}$$

Eq. 7.9 describes exponential growth: the concentration of B increases exponentially with the rate constant k from $B_0 = 0$ to $B_\infty = A_0$ (Figure 7.1). The amplitude of the reaction is A_0.

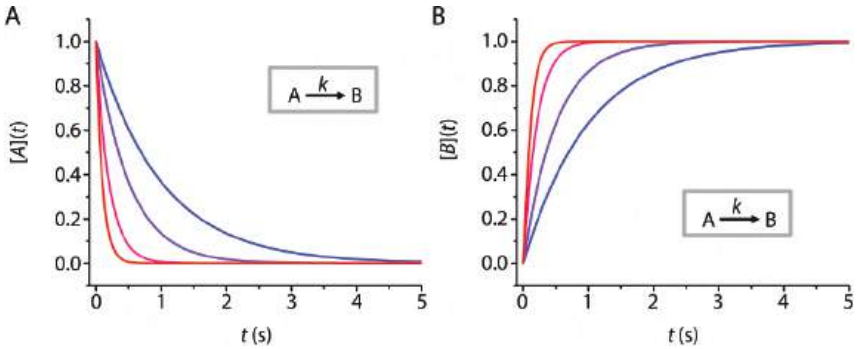


FIGURE 7.1: A first-order reaction A→B. A: The concentration of A decays exponentially from A_0 to zero. The rate constant determines how rapidly A is consumed. B: The concentration of the product B increases exponentially from zero to $B_\infty = A_0$. The set of plots corresponds to $k = 1\ s^{-1}$ (blue), $k = 2\ s^{-1}$ (violet), $k = 5\ s^{-1}$ (pink), and $k = 10\ s^{-1}$ (red).

To determine the rate constant for the reaction from A to products experimentally, we need to measure the concentration $[A]$ as a function of time, either continuously or discontinuously (see Chapter 28 for kinetic methods). The rate constant k for the reaction can then be determined by non-linear regression using eq. 7.7 (Figure 7.2, Box 7.1).

BOX 7.1: DETERMINATION OF RATE CONSTANTS FROM EXPERIMENTAL DATA.

To determine a rate constant for a chemical reaction, we monitor the concentration(s) of reactant(s) or product(s) as a function of time, either continuously or discontinuously (Figure 7.2). We then describe the experimental data points by a single exponential function, varying the amplitude and the rate constant until we obtain the optimal description of the data points. The rate constant that gives the best fit between curve and data is the rate constant k of the reaction.

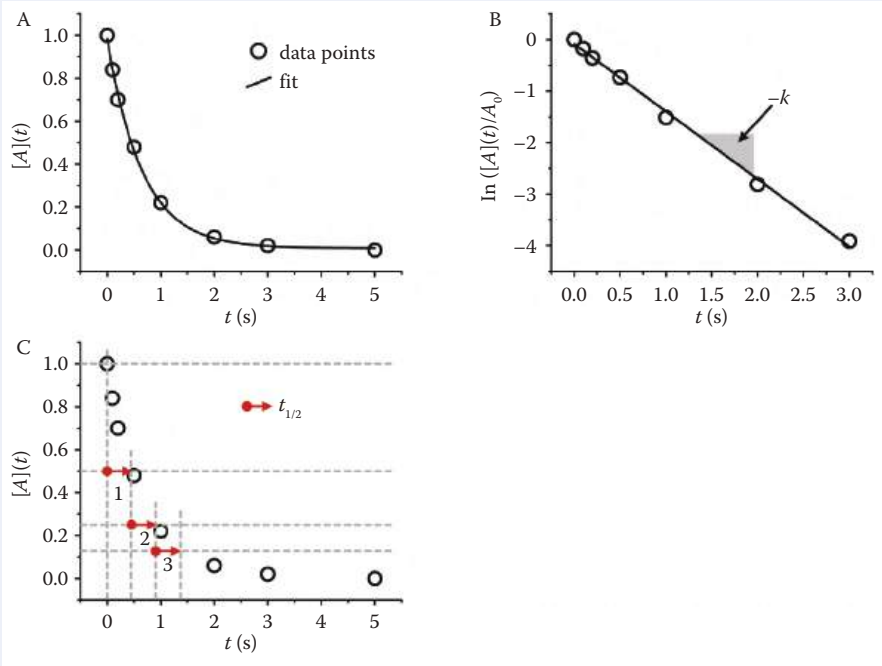


FIGURE 7.2: Experimental data and description by a first-order integrated rate law. A: Fitting curves to experimental data. Experimental data points for the decrease in concentration of the reactant are best described by an exponential decay (eq. 7.7) with the rate constant $k = 1.52\ s^{-1}$: the reaction is a first-order reaction with a rate constant $k = 1.52\ s^{-1}$. B: Linearization according to eq. 7.6 and determination of k from the slope gives $k = 1.31\ s^{-1}$. C: Determination of the reaction order and the rate constant from half-lives (red arrows) according to eq. 7.13. Three half-lives are approximately 1.4 s, corresponding to $k = 1.49\ s^{-1}$. The obtained rate constants differ because of different weighting (B) and approximation errors (C).

Alternatively, we can use the linear relationship between $\ln([A]/A_0)$ and the time t according to eq. 7.6 to determine the rate constant of a first-order reaction. We obtain k from the slope of a plot of $\ln([A]/A_0)$ as a function of t (Figure 7.2). However, transformation of the concentrations to the logarithm transforms the experimental errors associated with measuring the concentrations in a non-linear fashion. Therefore, it is better to directly plot the experimentally determined concentrations as a function of time, and to describe the experimental data points by a curve according to eq. 7.7 to extract k. In general, linearized plots of converted data values should be avoided due to the error transformation (see Figure 13.4). Finally, first-order rate constants can be calculated from half-lives (eq. 7.13, Figure 7.2).

Often, eq. 7.7 is written in terms of a time constant τ instead of the rate constant k. The time constant τ of an exponential function is the time t after which the exponential function has decayed by $1/e$, to $A(\tau) = A_0/e$. For eq. 7.7, this translates into

$$[A](\tau) = \frac{A_0}{e} = A_0 e^{-k\tau} \qquad \text{eq. 7.10}$$

Solving this equation for τ yields

$$\tau = \frac{1}{k} \qquad \text{eq. 7.11}$$

The time constant τ of the reaction has units of s, and is sometimes also called the *lifetime* of A. Fast reactions are characterized by high rate constants k and thus small lifetimes τ, whereas slow reactions have a small k and the lifetime τ of the reactant is high. The time constant or lifetime τ describes how long on average the molecules remain as A before they react to products.

In analogy to the time constant or lifetime τ, we can also define a *half-life* $t_{1/2}$, the time at which the starting concentration A_0 has halved. From

$$[A](t_{1/2}) = \frac{A_0}{2} = A_0 e^{-kt_{1/2}} \qquad \text{eq. 7.12}$$

we obtain the half-life $t_{1/2}$ as

$$t_{1/2} = \frac{\ln 2}{k} \approx \frac{0.7}{k} \qquad \text{eq. 7.13}$$

It is noteworthy that the half-life for a first-order reaction is independent of the starting concentration A_0 (Figure 7.3). As a consequence, the first half-life, *i.e.* the time for the concentration to be reduced from A_0 to $A_0/2$, is identical to the second half-life (from $A_0/2$ to $A_0/4$), the third half-life (from $A_0/4$ to $A_0/8$), and so on. Because of their constant half-lives, first-order reactions can be used as a reference clock. An example is dating of material by the radiocarbon method (Box 7.2).

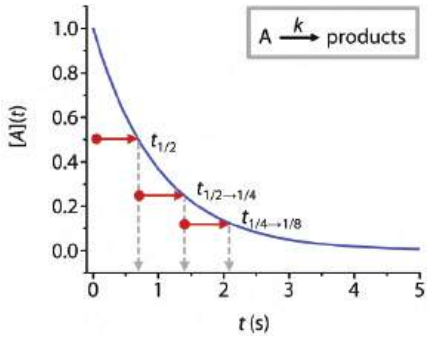


FIGURE 7.3: Half-lives in a first-order reaction. For a first-order reaction, the half-life $t_{1/2}$ is independent of the starting concentration A_0. Therefore, the first half-life $t_{1/2}$, the time for the concentration to be reduced from A_0 to $A_0/2$, is identical to the second half-life ($t_{1/2\rightarrow1/4}$, from $A_0/2$ to $A_0/4$), the third half-life ($t_{1/4\rightarrow1/8}$, from $A_0/4$ to $A_0/8$), and so on.

BOX 7.2: A FIRST-ORDER REACTION: RADIOACTIVE DECAY AND RADIOCARBON DATING.

Radioactive decay is a prominent example of a first-order reaction. At any given time, the velocity v of the decay reaction is proportional to the number N of radioactive molecules present, with the rate constant k as the proportionality constant:

$$v = -\frac{dN}{dt} = kN \qquad \text{eq. 7.14}$$

Eq. 7.14 is analogous to eq. 7.3, and by rearranging and integrating, we obtain

$$N(t) = N_0 e^{-kt} \qquad \text{eq. 7.15}$$

This integrated rate law is formally identical to the first-order rate law in eq. 7.7. Radioactive isotopes are characterized by their half-lives $t_{1/2}$ (eq. 7.13, Table 7.1).

TABLE 7.1: Half-lives of radioisotopes.

isotope	half-life (approx.)	use
^{14}C	5730 a	radiocarbon dating, radiotracer in biochemistry
^{3}H	12 a	radiotracer in biochemistry
^{35}S	87 d	radiotracer in biochemistry
^{125}I	59 d	nuclear medicine: determination of bone density
^{32}P	14 d	radiotracer in biochemistry
^{131}I	8 d	nuclear medicine: diagnostics of thyroid function
^{18}F	1.8 h	nuclear medicine: radiotracer in positron emission tomography

Carbon occurs in nature as three different *isotopes*, ^{12}C (> 98.8%), ^{13}C (~1%), and ^{14}C (< 10^{-10} %). ^{12}C and ^{13}C are stable isotopes, whereas ^{14}C radioactively decays to ^{14}N with a half-life of 5730 years. Living organisms maintain a constant level of ^{14}C in their bodies: although ^{14}C decays to ^{14}N, the ^{14}C levels are replenished through exchange with the surroundings. Once the organisms die, metabolism stops, and ^{14}C is not replenished anymore. The ^{14}C level then decreases because of its radioactive decay, following a first-order rate law:

$$\left[{}^{14}C\right](t) = \left[{}^{14}C\right]_0 e^{-kt} = \left[{}^{14}C\right]_0 e^{-\frac{\ln 2}{t_{1/2}} \cdot t} \qquad \text{eq. 7.16}$$

In eq. 7.16, $t = 0$ is the time when metabolism stopped. $[^{14}C]_0$ is the constant level of ^{14}C in living organisms (10^{-10} %) that was present at $t = 0$, $t_{1/2}$ is the half-life of ^{14}C (5730 years), and $[^{14}C](t)$ is the concentration of ^{14}C at the time t after death of the organism. By determining the ^{14}C level of organic material that contains carbon, the time t that has elapsed since the organism died can be calculated. Radiocarbon dating is a standard method in archaeology, and it has been used to date wood, seeds, pollen, corals, shells, bones, hair, peat, soil, fabrics, paper, paint, and pottery. After ten half-lives, the amount of ^{14}C is below the levels that can be quantified, limiting the age determination to < 57300 years.

We now derive the rate law for a bimolecular reaction. First, we consider a reaction of two (identical) molecules of A to products (Scheme 7.2):

$$A + A \xrightarrow{k} \text{products}$$ scheme 7.2

The association of two protein monomers to a dimer is an example of such a type of reaction (see Box 7.4). The definition of the reaction velocity v as the change in concentration of A with time is

$$v = -\frac{1}{2} \cdot \frac{d[A]}{dt}$$ eq. 7.17

The factor ½ in eq. 7.17 takes into account that two molecules of A react to products. Per product formed, the concentration of A thus decays twice as much compared to a unimolecular reaction (Scheme 7.1). When we express the reaction velocity for the bimolecular reaction as a function of the rate constant and the actual concentration of A, we also have to take into account that two molecules of A react with each other. The concentration of A therefore appears twice as a factor in the rate law, and the reaction velocity depends on the square of the concentration of A: the reaction is a second-order reaction.

$$v = -\frac{1}{2} \cdot \frac{d[A]}{dt} = k[A][A] = k[A]^2$$ eq. 7.18

To obtain the reaction velocity v in units of M s^{-1}, the product on the right-hand side of eq. 7.18 must also yield M s^{-1}. With the two concentration terms (units M), we can derive the unit M^{-1} s^{-1} for the rate constant k of a second-order reaction.

In order to integrate the rate law, we again separate the variables in eq. 7.18:

$$\frac{d[A]}{[A]^2} = -2k\,dt$$ eq. 7.19

By integration from A_0 to $[A](t)$, and from 0 to t, we obtain

$$\int_{A_0}^{[A](t)} \frac{d[A]}{[A]^2} = -2k \int_0^t dt$$ eq. 7.20

and

$$-\frac{1}{[A](t)} + \frac{1}{A_0} = -2kt$$ eq. 7.21

which can be rearranged to:

$$\frac{1}{[A](t)} = 2kt + \frac{1}{A_0}$$ eq. 7.22

The integrated rate law in eq. 7.22 reveals a linear dependence between the inverse of $[A](t)$ and t. From a linear plot of $1/[A]$ as a function of t, we can determine the second-order rate constant k from the slope ($2k$). Alternatively, we can rearrange to:

$$[A](t) = \frac{A_0}{1 + 2ktA_0}$$ eq. 7.23

and can directly plot $[A]$ as a function of time (Figure 7.4). The concentration of A initially decays more rapidly compared to a first-order reaction, but then approaches zero more slowly than it does for a first-order reaction (Figure 7.4). In other words, we can directly distinguish between a first-order reaction (Scheme 7.1) and a second-order reaction (Scheme 7.2) from the shape of the $[A](t)$ curves (Box 7.3).

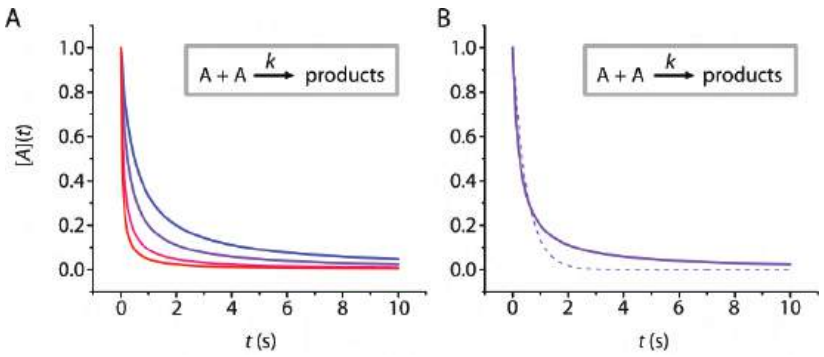


FIGURE 7.4: A second-order reaction: A+A → products. A: Concentration of A (in μM) as a function of time for a second-order rate constant of $k = 1\ \mu M^{-1}\ s^{-1}$ (blue), $k = 2\ \mu M^{-1}\ s^{-1}$ (violet), $k = 5\ \mu M^{-1}\ s^{-1}$ (pink), and $k = 10\ \mu M^{-1}\ s^{-1}$ (red). B: Comparison of $[A](t)$ for a second-order reaction (solid line) with $k = 2\ \mu M^{-1}\ s^{-1}$ and a first-order reaction (dashed line) with the rate constants $k = 2\ s^{-1}$. The concentration of A decays more rapidly in the beginning than for the corresponding first-order reaction with the same rate constant, but levels off more slowly.

BOX 7.3: DISTINCTION BETWEEN FIRST- AND SECOND-ORDER REACTIONS.

The dependence of the concentration of the reactant as a function of time differs for first- and second-order reactions. To distinguish between first- and second-order reactions, we plot the experimentally determined concentration of the reactant as a function of time, and try to describe the data with the integrated rate laws for first- and second-order reactions (Figure 7.5). The correct integrated rate law describes the experimental data, while the other rate law yields a curve that systematically deviates from the data.

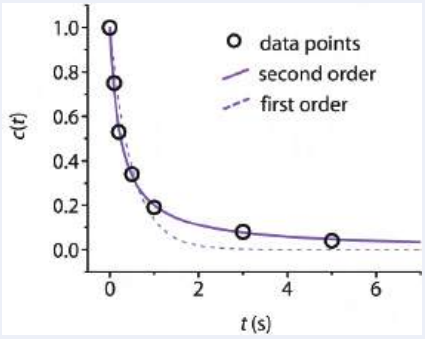


FIGURE 7.5: Distinction between first- and second-order reactions. Experimentally determined concentration of the reactant as a function of time (black circles) and best fit of curves for a first-order reaction (dotted line, eq. 7.7, $k = 2\ s^{-1}$) and a second-order reaction (solid line, eq. 7.23, $k = 2\ s^{-1}$). The single exponential curve deviates systematically from the data, indicating that the reaction is not a first-order reaction.

We can calculate the half-life of A for the second-order reaction as

$$[A](t_{1/2}) = \frac{A_0}{2} = \frac{A_0}{1 + 2kt_{1/2}A_0} \qquad \text{eq. 7.24}$$

and obtain

$$t_{1/2} = \frac{1}{2kA_0} \qquad \text{eq. 7.25}$$

Note that in contrast to the first-order reaction, where the half-life is independent of the initial concentration of A, the half-life for the second-order reaction now depends on the initial concentration of A. As a consequence, the first half-life, from A_0 to $A_0/2$, is different from the second half-life, the time passing between $A_0/2$ and $A_0/4$. By comparing the half-life of a reaction at different starting concentrations we can therefore also distinguish between first- and second-order reactions. From

$$[A](t_{1/4}) = \frac{A_0}{4} = \frac{A_0}{1 + 2kt_{1/4}A_0} \qquad \text{eq. 7.26}$$

we obtain $t_{1/4}$ as

$$t_{1/4} = \frac{3}{2kA_0} \qquad \text{eq. 7.27}$$

The second half-life is the time that has passed from half to a quarter of the starting concentration of A and can be calculated as the difference between $t_{1/4}$ and $t_{1/2}$:

$$t_{1/4} - t_{1/2} = \frac{2}{2kA_0} \qquad \text{eq. 7.28}$$

From a comparison of eq. 7.25 and eq. 7.28 we see that the second half-life thus is twice the first half-life. The third half-life (from $A_0/4$ to $A_0/8$) is twice the second half life, and so on (Figure 7.6).

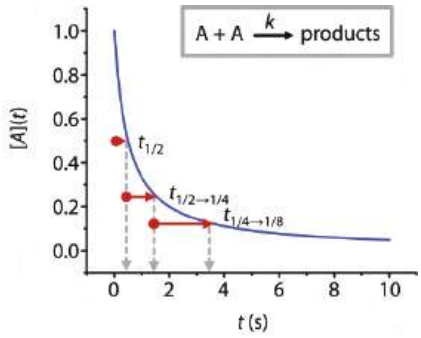


FIGURE 7.6: Half-lives in a second-order reaction. For a second-order reaction, the half-life $t_{1/2}$ depends on the starting concentration A_0. As a consequence, the second half-life, the time passing between $A_0/2$ to $A_0/4$, is twice as long as the first half-life (from A_0 to $A_0/2$), the third half-life (from $A_0/4$ to $A_0/8$) is twice as long as the second, and so on.

A different type of a bimolecular reaction is a reaction of A with a different molecule B to products with the rate constant k (Scheme 7.3):

$$\mathrm{A + B} \xrightarrow{k} \text{products} \qquad \text{scheme 7.3}$$

This scheme describes any association of two different molecules, and thus any binary complex formation (Box 7.4). The corresponding rate law is

$$v = -\frac{\mathrm{d}[A]}{\mathrm{d}t} = -\frac{\mathrm{d}[B]}{\mathrm{d}t} = k[A][B] \qquad \text{eq. 7.29}$$

The reaction is of first order with respect to A and to B, but of second order overall. The unit of k is $M^{-1}\ s^{-1}$. We can now separate the variables

$$-\frac{\mathrm{d}[A]}{[A][B]} = k\,\mathrm{d}t \qquad \text{eq. 7.30}$$

In eq. 7.30, both [*A*] and [*B*] depend on the time *t*. To obtain a form of this equation that we can integrate, we express the current concentration of A as A_0 minus the concentration x that has already reacted to products. As one molecule A reacts with one molecule of B, we can express the concentration of B as B_0 minus the same concentration x. x thus represents the product formed, and the rate law for product formation is

$$\frac{\mathrm{d}x}{\mathrm{d}t} = k[A][B] \qquad \text{eq. 7.31}$$

We thereby reduce the equation to a single integration variable x and its change $\mathrm{d}x = -\mathrm{d}A$. Eq. 7.30 then becomes

$$\frac{\mathrm{d}x}{(A_0 - x)(B_0 - x)} = k\,\mathrm{d}t \qquad \text{eq. 7.32}$$

which we can now integrate from 0 to $x(t)$ and 0 to t:

$$\int_0^{x(t)} \frac{\mathrm{d}x}{(A_0 - x)(B_0 - x)} = \int_0^t k\,\mathrm{d}t \qquad \text{eq. 7.33}$$

The right-hand side of eq. 7.33 can be integrated to kt. To integrate the left-hand side, we need to express the fraction by partial fractions (see Section 31.7) and then integrate the partial fractions individually by substitution (see Section 31.6). We can express the fraction on the left-hand side as

$$\frac{1}{(A_0 - x)(B_0 - x)} = \frac{1}{(A_0 - B_0)} \cdot \left(\frac{1}{(B_0 - x)} - \frac{1}{(A_0 - x)} \right) \qquad \text{eq. 7.34}$$

We now have to evaluate the integral

$$\int_0^{x(t)} \frac{1}{(A_0 - B_0)} \cdot \left(\frac{1}{(B_0 - x)} - \frac{1}{(A_0 - x)} \right) \mathrm{d}x = kt \qquad \text{eq. 7.35}$$

$1/(A_0 - B_0)$ is a constant factor that remains on integration. The two fractions in the parentheses can be integrated separately, giving

$$\frac{1}{(A_0 - B_0)} \cdot \left(\left[-\ln(B_0 - x) \right]_0^{x(t)} - \left[-\ln(A_0 - x) \right]_0^{x(t)} \right) + \text{const.} = kt \qquad \text{eq. 7.36}$$

or

$$\frac{1}{(A_0 - B_0)} \cdot \left(-\ln(B_0 - x(t)) + \ln B_0 - \left(-\ln(A_0 - x(t)) + \ln A_0 \right) \right) + \text{const.} = kt \qquad \text{eq. 7.37}$$

By rearranging the logarithmic terms, we obtain

$$\frac{1}{(A_0 - B_0)} \cdot \left(\ln \frac{A_0 - x(t)}{B_0 - x(t)} - \ln \frac{A_0}{B_0} \right) + \text{const.} = kt \qquad \text{eq. 7.38}$$

The boundary condition $x(0) = 0$ is fulfilled with an integration constant (const.) of zero. Eq. 7.38 can then be solved for $x(t)$ through multiplication by $(A_0 - B_0)$,

$$\left(\ln \frac{A_0 - x(t)}{B_0 - x(t)} - \ln \frac{A_0}{B_0} \right) = k(A_0 - B_0)t \qquad \text{eq. 7.39}$$

combination of the logarithmic terms on the left-hand side to

$$\ln\frac{\left(A_0-x(t)\right)B_0}{\left(B_0-x(t)\right)A_0}=k\left(A_0-B_0\right)t \qquad \text{eq. 7.40}$$

application of the exponential function on both sides,

$$\frac{\left(A_0-x(t)\right)B_0}{\left(B_0-x(t)\right)A_0}=\exp\left(k\left(A_0-B_0\right)t\right) \qquad \text{eq. 7.41}$$

multiplication by the denominator of the left-hand side,

$$A_0-x(t)=\left(B_0-x(t)\right)\frac{A_0}{B_0}\exp\left(k\left(A_0-B_0\right)t\right) \qquad \text{eq. 7.42}$$

isolation of the terms containing $x(t)$,

$$-x(t)+x(t)\frac{A_0}{B_0}\exp\left(k\left(A_0-B_0\right)t\right)=-A_0+A_0\exp\left(k\left(A_0-B_0\right)t\right) \qquad \text{eq. 7.43}$$

and

$$x(t)\left(\frac{A_0}{B_0}\exp\left(k\left(A_0-B_0\right)t\right)-1\right)=-A_0+A_0\exp\left(k\left(A_0-B_0\right)t\right) \qquad \text{eq. 7.44}$$

division by the factor following $x(t)$,

$$x(t)=\frac{-A_0+A_0\exp\left(k\left(A_0-B_0\right)t\right)}{\left(\frac{A_0}{B_0}\exp\left(k\left(A_0-B_0\right)t\right)-1\right)} \qquad \text{eq. 7.45}$$

and rearrangement to the integrated rate law

$$x(t)=\frac{A_0-A_0e^{k\left(A_0-B_0\right)t}}{1-\frac{A_0}{B_0}e^{k\left(A_0-B_0\right)t}} \qquad \text{eq. 7.46}$$

$[A](t)$ and $[B](t)$ can be calculated as A_0–$x(t)$ and B_0–$x(t)$ from mass conservation, which gives

$$[A](t)=A_0-\frac{A_0-A_0e^{k\left(A_0-B_0\right)t}}{1-\frac{A_0}{B_0}e^{k\left(A_0-B_0\right)t}} \qquad \text{eq. 7.47}$$

and

$$[B](t)=B_0-\frac{A_0-A_0e^{k\left(A_0-B_0\right)t}}{1-\frac{A_0}{B_0}e^{k\left(A_0-B_0\right)t}} \qquad \text{eq. 7.48}$$

The concentrations $[A](t)$, $[B](t)$, and $x(t)$ are depicted in Figure 7.7.

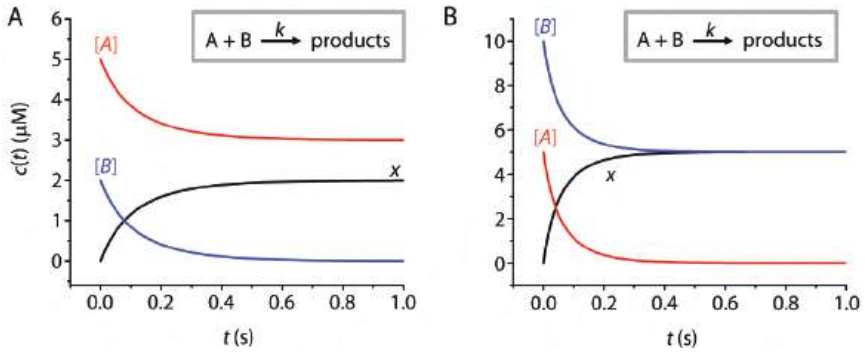


FIGURE 7.7: Second-order reaction A+B → products. $[A](t)$, $[B](t)$, and $x(t)$ for a bimolecular reaction of A and B to products. A: $A_0 = 5$ μM, $B_0 = 2$ μM, and $k = 2$ s^{-1}. B: $A_0 = 5$ μM, $B_0 = 10$ μM, and $k = 2$ s^{-1}.

Note that eq. 7.48 and all derived forms are not defined for $A_0 = B_0$. For this special case, we have derived the rate law before: for $A_0 = B_0$, and a reaction where one molecule A reacts with one molecule of B, the concentrations of A and B are equal at any given point in time, $[A](t) = [B](t)$. Therefore, the rate law can be written as

$$-\frac{\mathrm{d}[A]}{\mathrm{d}t} = -\frac{\mathrm{d}[B]}{\mathrm{d}t} = k[A][B] = k[A]^2 \qquad \text{eq. 7.49}$$

which is identical to eq. 7.18 for the second-order reaction A+A → products, leading to the identical equation for $[A](t)$ (eq. 7.23).

BOX 7.4: SECOND-ORDER REACTIONS: DIMERIZATION AND LIGAND BINDING.

Many proteins need to associate to dimers to perform their biological functions. Examples are glutathione-S-transferase, alcohol dehydrogenase, transcription factors, and type II DNA restriction enzymes. After biosynthesis and folding of the monomeric forms, two monomers form a homodimer in a second-order reaction according to Scheme 7.2. The formation of heterodimers by two different subunits is a second-order reaction following Scheme 7.3. In fact, any binary complex formation is described by Scheme 7.3, including the association of two complementary strands of DNA or RNA, protein-protein interactions such as the complex formation of Ras-like small G-proteins with their effectors, protein-ligand interactions such as binding of a substrate, inhibitor or activator molecule to an enzyme, as well as protein-nucleic acid interactions.

We can apply the same principles we have used for unimolecular and bimolecular reactions to formulate rate laws and derive integrated rate laws, expressions for the concentrations of reaction partners as a function of time, for trimolecular reactions and reactions of higher molecularity. However, trimolecular reactions and reactions of higher molecularity are very rare (see Section 15.3), and will therefore not be treated.

QUESTIONS

7.1 You measure the concentration of a reactant as a function of time and obtain the following dataset:

t (s)	$c(t)$ (µM)
0	10
1	7.1
2	5.0
3	3.5
4	2.5
5	1.7
6	1.3
8	0.6
10	0.3

What are the reaction order and the rate constant?

7.2 A reactant shows the following decrease in concentration with reaction time:

t (s)	$c(t)$ (µM)
0	1
1	0.74
2	0.59
3	0.49
4	0.42
5	0.37
6	0.32
8	0.27
10	0.22

Determine the reaction order and estimate the rate constant.

7.3 An old tree trunk is discovered and submitted to radiocarbon dating. The ^{14}C percentage in the analyzed samples is 10^{-12}%. What is the age of the tree? $[^{14}C]_0 = 10^{-10}$ %, $t_{½} = 5730$ a.

7.4 You determine the concentration of a reactant that reacts on its own at time $t = 3.6$ s as $c(t) = 0.27$ mM. The starting concentration was 10 mM. Does the reaction follow the rate law for first- and second-order reactions? What is the rate constant?

7.5 The reactants A and B react to products in an irreversible reaction. What is the concentration of A and B at $t = 5$ s for (1) $A_0 = 1$ µM, $B_0 = 10$ µM, (2) $A_0 = 10$ µM, $B_0 = 1$ µM, (3) $A_0 = 6$ µM, $B_0 = 5$ µM, and for (4) $A_0 = 20$ µM, $B_0 = 29$ µM? $k = 0.2$ µM^{-1} s^{-1}.

7.6 What is the mean rate constant for a reaction in which a compound decays to products in a double-exponential decay, with individual rate constants k_1 and k_2 and amplitudes of f_1 and f_2 for the two reaction phases?

7.7 Rate constants for reactions are often determined from initial velocities, *i.e.* from the linear increase of reaction velocities shortly after the reaction was started. Show that this approach is valid for a first-order reaction with the rate constant k, both for the decrease of the concentration of reactant and for the increase in product concentration.

7.8 What is the initial velocity of a reaction described by a double-exponential rate law?

7.9 Derive the integrated rate law for the trimolecular reaction 3 A → products. The rate constant is k. What is a suitable plot to determine the rate constant by linear regression?

8 Reaction types

In Chapter 7, we have calculated integrated rate laws for isolated reactions. In the following, we will derive rate laws for reaction schemes comprising two reactions: reversible reactions, parallel reactions, and sequential reactions. These reaction types are important elements of more complex reaction schemes. Their treatment will enable us to analyze and quantitatively understand complex biochemical reaction pathways and networks.

8.1 REVERSIBLE REACTIONS

We first look at a *reversible reaction* in which A is converted to B in a forward reaction with the rate constant k_1, and B is converted to A in a reverse reaction with the rate constant k_{-1} (Scheme 8.1):

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B$$ scheme 8.1

Examples of reversible reactions are isomerizations, such as the interconversion of dihydroxyacetone phosphate and glyceraldehyde-3-phosphate in glycolysis, protein folding according to the two-state model (see Section 18.5.4), or conformational changes of proteins, *e.g.* between open and closed states (see Section 4.3). We can express the rate law for a reversible reaction in terms of the forward and reverse reactions as

$$v = \frac{dc}{dt} = -\frac{d[A]}{dt} = \frac{d[B]}{dt}$$ eq. 8.1

The change in concentration of A is brought about by the forward reaction that leads to a consumption of A, and thus a decrease in its concentration. The concentration change over time is the product of the rate constant k_1 and the actual concentration of A. This term appears with a negative sign to take into account the decrease of $[A]$ caused by this process. The reverse reaction, from B to A, leads to an increase in the concentration of A. Its velocity is the product of the rate constant k_{-1} and the current concentration of B, and this term enters the balance with a positive sign due to the increase in $[A]$ (see also Box 8.2). The change in $[A]$ over time is thus

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k_1[A] + k_{-1}[B] \qquad \text{eq. 8.2}$$

We know from the law of mass conservation that the total concentration of A (A_0) that was present at $t = 0$ will either remain as A or will have been converted to B. When the reaction starts from A, *i.e.* the starting concentration of B is zero, eq. 8.3

$$[A] + [B] = A_0 \qquad \text{eq. 8.3}$$

has to be fulfilled at any given time. We can now solve eq. 8.3 for [*B*], and substitute [*B*] in eq. 8.2, and obtain

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k_1[A] + k_{-1}\left(A_0 - [A]\right) = -\left(k_1 + k_{-1}\right)[A] + k_{-1}A_0 \qquad \text{eq. 8.4}$$

Separating the variables gives

$$\frac{\mathrm{d}[A]}{-\left(k_1 + k_{-1}\right)[A] + k_{-1}A_0} = \mathrm{d}t \qquad \text{eq. 8.5}$$

We can now integrate eq. 8.5 by substitution: we substitute the denominator for u, and obtain

$$\frac{\mathrm{d}u}{\mathrm{d}[A]} = -\left(k_1 + k_{-1}\right) \quad \text{or} \quad \mathrm{d}[A] = \frac{\mathrm{d}u}{-\left(k_1 + k_{-1}\right)} \qquad \text{eq. 8.6}$$

The integral then becomes

$$\frac{1}{-\left(k_1 + k_{-1}\right)} \cdot \int \frac{1}{u}\mathrm{d}u = \int_0^t \mathrm{d}t \qquad \text{eq. 8.7}$$

or

$$\int \frac{1}{u}\mathrm{d}u = \int_0^t \left(-\left(k_1 + k_{-1}\right)\right)\mathrm{d}t \qquad \text{eq. 8.8}$$

and can be evaluated as

$$\ln u = -\left(k_1 + k_{-1}\right)t \qquad \text{eq. 8.9}$$

Substituting back u for the denominator in eq. 8.5 yields

$$\ln u = \left[\ln\left(-\left(k_1 + k_{-1}\right)[A] + k_{-1}A_0\right)\right]_{A_0}^{[A](t)} = -\left(k_1 + k_{-1}\right)t \qquad \text{eq. 8.10}$$

We can now evaluate the integral on the left from A_0 to $[A](t)$ to

$$\ln\left(-\left(k_1 + k_{-1}\right)[A](t) + k_{-1}A_0\right) - \ln\left(-\left(k_1 + k_{-1}\right)A_0 + k_{-1}A_0\right) = -\left(k_1 + k_{-1}\right)t \qquad \text{eq. 8.11}$$

Eq. 8.11 can be solved for $[A](t)$ by combining the logarithmic terms on the left-hand side,

$$\ln\frac{\left(-\left(k_1+k_{-1}\right)[A](t)+k_{-1}A_0\right)}{\left(-\left(k_1+k_{-1}\right)A_0+k_{-1}A_0\right)}=-\left(k_1+k_{-1}\right)t \qquad \text{eq. 8.12}$$

applying the exponential function on both sides,

$$\frac{-\left(k_1+k_{-1}\right)[A](t)+k_{-1}A_0}{-\left(k_1+k_{-1}\right)A_0+k_{-1}A_0}=e^{-(k_1+k_{-1})t} \qquad \text{eq. 8.13}$$

multiplying by the denominator of the left-hand side,

$$-\left(k_1+k_{-1}\right)[A](t)+k_{-1}A_0=\left(-\left(k_1+k_{-1}\right)A_0+k_{-1}A_0\right)\cdot e^{-(k_1+k_{-1})t} \qquad \text{eq. 8.14}$$

dividing by the factor in front of $[A](t)$,

$$[A](t)=\frac{\left(-\left(k_1+k_{-1}\right)A_0+k_{-1}A_0\right)\cdot e^{-(k_1+k_{-1})t}-k_{-1}A_0}{-\left(k_1+k_{-1}\right)} \qquad \text{eq. 8.15}$$

and by separating the fractions on the right-hand side, which gives

$$[A](t)=\frac{k_1A_0}{k_1+k_{-1}}\cdot e^{-(k_1+k_{-1})t}+\frac{k_{-1}A_0}{k_1+k_{-1}} \qquad \text{eq. 8.16}$$

or, by factoring out $A_0/(k_1+k_{-1})$:

$$[A](t)=\frac{A_0}{k_1+k_{-1}}\left(k_{-1}+k_1e^{-(k_1+k_{-1})t}\right) \qquad \text{eq. 8.17}$$

Eq. 8.17 contains a constant term and an exponential function that describes the exponential decay of $[A]$ as a function of time. Note that instead of the rate constant k in the exponent, as in the irreversible reaction (eq. 7.48), now the sum of the rate constants for the forward and reverse reaction, $k_1 + k_{-1}$, appears as an *apparent rate constant* k_{app}. The rate constants k_1 and k_{-1} are the *microscopic rate constants*.

We can calculate $[B](t)$ as a function of time from the mass conservation as $A_0-[A](t)$:

$$[B](t)=\frac{A_0\cdot k_1}{k_1+k_{-1}}\left(1-e^{-(k_1+k_{-1})t}\right) \qquad \text{eq. 8.18}$$

Again, the apparent rate constant for the appearance of B is the sum of the individual rate constants k_1 and k_{-1}. Thus, if we follow the disappearance of A or the appearance of B over time by a suitable spectroscopic signal (see Chapters 21 and 28) and describe the signal as a function of time by a single exponential, we can only determine the sum of these rate constants. To extract values for both k_1 and k_{-1}, we need to know the absolute amplitudes, *i.e.* we need to be able to convert our signal into concentrations of A and B.

We know that if a reaction occurs in both directions, an equilibrium between A and B will be established. The rate constant with which the equilibrium is approached is the same apparent rate constant for the consumption of A and the production of B (Figure 8.1). The apparent rate constant is independent of the starting conditions: no matter if we start from A or B, the equilibrium concentrations will be reached with $k_{app} = k_1 + k_{-1}$. In fact, although we have not derived it here, this is also true when we start from any non-equilibrium mixture of A and B.

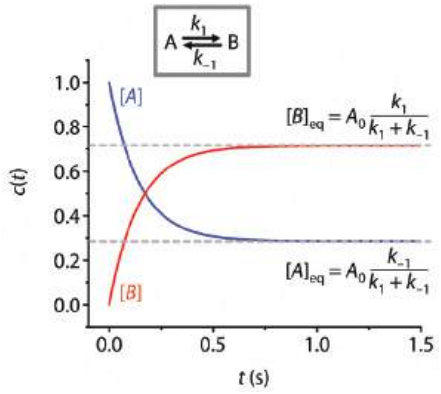


FIGURE 8.1: A reversible reaction. Concentrations $c(t)$ of A (blue) and B (red) as a function of time for rate constants $k_1 = 5\ s^{-1}$ and $k_{-1} = 2\ s^{-1}$. A decays exponentially to its equilibrium concentration $[A]_{eq}$ with an apparent rate constant of $k_{app} = k_1 + k_{-1} = 7\ s^{-1}$, and B grows exponentially to its equilibrium concentration $[B]_{eq}$ with the same apparent rate constant. The equilibrium constant $K = [B]_{eq}/[A]_{eq}$ is the ratio of the two rate constants k_1 and k_{-1}.

What can we say about the equilibrium concentrations $[A]_{eq}$ and $[B]_{eq}$, once the equilibrium has been established? To answer this question, we have to determine the concentrations of A and B for $t \to \infty$. In eq. 8.17 and eq. 8.18, the exponential term will become zero for $t \to \infty$, leaving us with the constant term

$$[A]_{eq} = \frac{k_{-1}A_0}{k_1 + k_{-1}} \qquad \text{eq. 8.19}$$

The concentration of B in equilibrium can be calculated from the mass conservation, or directly from the limit of eq. 8.18 for $t \to \infty$. In both cases, we obtain

$$[B]_{eq} = A_0 - [A]_{eq} = \frac{k_1 A_0}{k_1 + k_{-1}} \qquad \text{eq. 8.20}$$

We can now calculate the equilibrium constant K from the ratio of the equilibrium concentrations $[A]_{eq}$ and $[B]_{eq}$ as

$$K = \frac{[B]_{eq}}{[A]_{eq}} = \frac{\dfrac{k_1 A_0}{k_1 + k_{-1}}}{\dfrac{k_{-1}A_0}{k_1 + k_{-1}}} = \frac{k_1}{k_{-1}} \qquad \text{eq. 8.21}$$

Eq. 8.21 illustrates the relationship between thermodynamic data (equilibrium constant) and kinetic data (rate constants). The position of the equilibrium is the result of a *kinetic competition* of forward and reverse reactions: if both reactions occur with the same rate constant, K equals unity. If k_1 is *e.g.* 10-fold higher than k_{-1}, then there will be 10 times the amount of product compared to reactant, and *vice versa*.

Note that the rate constants k_1 and k_{-1} have different units if we consider the reversible association reaction of two molecules A and B to a complex AB (Scheme 8.2).

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AB \qquad \text{scheme 8.2}$$

In this case, k_1 is a second-order rate constant for a bimolecular reaction, with units $M^{-1}\ s^{-1}$, and k_{-1} is the first-order rate constant for a unimolecular reaction, with units s^{-1}. Nevertheless, the equilibrium constants K_a for complex formation or K_d for complex dissociation can still be expressed by the

ratio of the rate constants. The equilibrium constant for complex formation, K_a, in M^{-1} (see Section 2.2, Box 2.2) is

$$K_a = \frac{[AB]_{eq}}{[A]_{eq}[B]_{eq}} = \frac{k_1}{k_{-1}} \qquad \text{eq. 8.22}$$

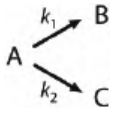
The dissociation constant K_d of the complex AB is the inverse of the association constant, and thus

$$K_d = \frac{[A]_{eq}[B]_{eq}}{[AB]_{eq}} = \frac{k_{-1}}{k_1} \qquad \text{eq. 8.23}$$

with units M. (Bio)chemical equilibria are often discussed in terms of K_d rather than K_a because of the more convenient units (see Section 2.2).

8.2 PARALLEL REACTIONS

We now consider a reaction scheme where A can react either to B or to C in two *parallel reactions* (Scheme 8.3). There are numerous examples of this reaction scheme in metabolism, where many reaction intermediates can enter two or more alternative pathways (see Figure 8.3).



scheme 8.3

The conversion of A to B is characterized by a rate constant k_1, and the rate constant for the conversion of A to C occurs with rate constant k_2. We can write the balance for the change in $[A]$ over time by summing the contributions from reactions consuming A:

$$\frac{d[A]}{dt} = -(k_1 + k_2) \cdot [A] \qquad \text{eq. 8.24}$$

B is produced from A, and the rate law is

$$\frac{d[B]}{dt} = k_1 \cdot [A] \qquad \text{eq. 8.25}$$

Similarly, C is produced from A according to the rate law

$$\frac{d[C]}{dt} = k_2 \cdot [A] \qquad \text{eq. 8.26}$$

By integration of eq. 8.24 from A_0 to $[A](t)$, we obtain the time dependence of A as

$$[A](t) = A_0 e^{-(k_1+k_2)t} \qquad \text{eq. 8.27}$$

Thus, A disappears with an apparent rate constant that is the sum of the individual rate constants, $k_{app} = k_1 + k_2$. We can substitute eq. 8.27 into eq. 8.25

$$\frac{d[B]}{dt} = k_1 \cdot A_0 e^{-(k_1+k_2)t} \qquad \text{eq. 8.28}$$

and, after separation of variables, obtain

$$d[B] = k_1 \cdot A_0 e^{-(k_1+k_2)t} dt \qquad \text{eq. 8.29}$$

which, on integration from B_0 to $[B](t)$ (left-hand side) and 0 to t (right-hand side)

$$\int_{B_0}^{[B](t)} \mathrm{d}[B] = \int_0^t k_1 \cdot A_0 e^{-(k_1+k_2)t} \mathrm{d}t \qquad \text{eq. 8.30}$$

gives

$$[B](t) - B_0 = k_1 A_0 \left(\frac{1}{-(k_1+k_2)} e^{-(k_1+k_2)t} + \frac{1}{k_1+k_2} \right) \qquad \text{eq. 8.31}$$

For the boundary condition $B_0 = 0$, we obtain for $[B](t)$

$$[B](t) = A_0 \frac{k_1}{k_1+k_2} \left(1 - e^{-(k_1+k_2)t}\right) \qquad \text{eq. 8.32}$$

Finally, we can calculate $[C](t)$ from mass conservation as $A_0 - [A](t) - [B](t)$ as

$$[C](t) = A_0 \frac{k_2}{k_1+k_2} \left(1 - e^{-(k_1+k_2)t}\right) \qquad \text{eq. 8.33}$$

B and C are thus also produced with an apparent rate constant $k_{app} = k_1 + k_2$ (Figure 8.2).

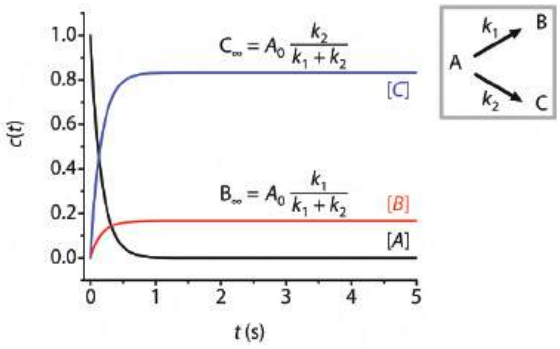


FIGURE 8.2: Parallel reactions. Concentrations $c(t)$ of A, B, and C as a function of time for a parallel reaction of A to B with rate constant k_1 (1 s^{-1}) and of A to C with rate constant k_2 (5 s^{-1}). A decays with an apparent rate constant of k_1+k_2, and the concentrations of B and C increase exponentially with the same apparent rate constant. The final ratios of B and C depend on the ratio of k_1 and k_2. The final concentration of B, B_∞, is A_0 multiplied with $k_1/(k_1+k_2)$, the final concentration of C, C_∞, equals A_0 times $k_2/(k_1+k_2)$. This situation is also called kinetic competition.

According to eq. 8.27–eq. 8.33, $[A](t)$, $[B](t)$, or $[C](t)$ from our kinetic experiment can all be described by a single exponential function, but the apparent rate constant determined from the data is the sum of the two individual rate constants k_1 and k_2. As discussed for the reversible reaction (Section 8.1), we can only determine the individual rate constants from $[B](t)$ or $[C](t)$ if we also know the absolute amplitude of the exponential function, *i.e.* B_∞ or C_∞.

What determines the final yield in B and C in such a reaction scheme? We can calculate the final concentrations of B and C for $t \to \infty$ from eq. 8.31 and eq. 8.33, and obtain the final ratio of B and C as

$$\frac{B_\infty}{C_\infty} = \frac{A_0 \dfrac{k_1}{k_1+k_2}}{A_0 \dfrac{k_2}{k_1+k_2}} = \frac{k_1}{k_2} \qquad \text{eq. 8.34}$$

Thus, the ratio of the rate constants k_1 and k_2 determines the ratio of the two products. Again, we have a kinetic competition between the two reactions.

We find numerous examples of parallel reactions in biochemistry (Figure 8.3). Every time a metabolic intermediate is a branch point and can be channeled into different pathways, the relative entry into the individual pathways (*partitioning*) is determined by kinetic competition. In glycolysis, for example, pyruvate can be further metabolized to acetyl-CoA, and then enter the citric acid cycle, or it can be used for the biosynthesis of alanine. Acetyl-CoA itself can either be channeled into the citric acid cycle or it can be used for the synthesis of fatty acids. Glyceraldehyde-3-phosphate is metabolized in the glycolytic pathway to 1,3-bisphosphoglycerate or used for the conversion of sugars in the pentose phosphate pathway. The overall reaction velocity of each reaction sequence after the branch point is determined by the slowest step of the sequence, the rate-limiting step (Chapter 10). The ratio of the rate constants for the rate-limiting step determines the fractions of the metabolic intermediate that will accumulate as the end products of the two alternative pathways.

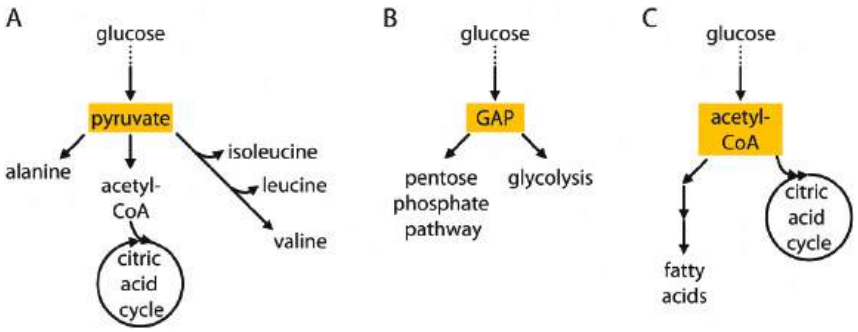


FIGURE 8.3: Metabolic branch points and parallel reactions. A: Pyruvate generated from glucose in glycolysis can be converted to alanine, channeled into the synthesis pathway for the amino acids isoleucine, leucine, and valine, or converted to acetyl-CoA that enters the citric acid cycle. B: Glyceraldehyde-3-phosphate (GAP) is either metabolized *via* the glycolysis pathway or enters the pentose phosphate pathway. C: Acetyl-CoA can either be channeled into the citric acid cycle or enter the synthesis pathway of fatty acids. The distribution over the different pathways depends on the rate constants of the respective processes (kinetic partitioning).

8.3 CONSECUTIVE REACTIONS

Finally, we consider a sequence of two reactions from A to B with k_1, and from B to C with k_2, according to Scheme 8.4:

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \qquad \text{scheme 8.4}$$

Virtually all metabolic pathways, such as the reaction sequence of glycolysis, or the synthesis of fatty acids or nucleobases, are *consecutive reactions*. The rate laws for A, B, and C for a reaction sequence A→B→C according to Scheme 8.4 are

$$\frac{d[A]}{dt} = -k_1 \cdot [A] \qquad \text{eq. 8.35}$$

and

$$\frac{d[B]}{dt} = k_1 \cdot [A] - k_2 \cdot [B] \qquad \text{eq. 8.36}$$

and

$$\frac{\mathrm{d}[C]}{\mathrm{d}t} = k_2 \cdot [B] \qquad \text{eq. 8.37}$$

We can integrate eq. 8.35 to obtain $[A](t)$ as

$$[A](t) = A_0 e^{-k_1 t} \qquad \text{eq. 8.38}$$

The concentration of A thus decays exponentially with the rate constant k_1, and shows identical behavior as in an isolated first-order reaction without subsequent reactions (eq. 7.7). The expression for $[A](t)$ may be substituted into eq. 8.36, which gives

$$\frac{\mathrm{d}[B]}{\mathrm{d}t} = k_1 \cdot A_0 e^{-k_1 t} - k_2 \cdot [B] \qquad \text{eq. 8.39}$$

or

$$\frac{\mathrm{d}[B]}{\mathrm{d}t} + k_2 \cdot [B] = k_1 \cdot A_0 e^{-k_1 t} \qquad \text{eq. 8.40}$$

Eq. 8.40 can then be integrated by multiplying both sides with $\exp(k_2 t)$ to give

$$\frac{\mathrm{d}[B]}{\mathrm{d}t} e^{k_2 t} + k_2 e^{k_2 t} \cdot [B] = k_1 \cdot A_0 e^{-k_1 t} e^{k_2 t} = k_1 \cdot A_0 e^{(k_2 - k_1)t} \qquad \text{eq. 8.41}$$

With

$$k_2 e^{k_2 t} = \frac{\mathrm{d}}{\mathrm{d}t} e^{k_2 t} \qquad \text{eq. 8.42}$$

we can substitute to

$$\frac{\mathrm{d}[B]}{\mathrm{d}t} e^{k_2 t} + \frac{\mathrm{d}}{\mathrm{d}t} e^{k_2 t} \cdot [B] = k_1 \cdot A_0 e^{(k_2 - k_1)t} \qquad \text{eq. 8.43}$$

The left-hand side of eq. 8.43 is of the form of the product law (Section 31.6), and equals $(uv)'$ with

$$u = e^{k_2 t} \qquad \text{eq. 8.44}$$

and

$$v = [B] \qquad \text{eq. 8.45}$$

which gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{k_2 t} \cdot [B]\right) = k_1 \cdot A_0 e^{(k_2 - k_1)t} \qquad \text{eq. 8.46}$$

By separating the integration variables, we obtain

$$\mathrm{d}\left(e^{k_2 t} \cdot [B]\right) = k_1 \cdot A_0 e^{(k_2 - k_1)t} \mathrm{d}t \qquad \text{eq. 8.47}$$

and can now integrate

$$\int_0^{B(t)} \mathrm{d}\left(e^{k_2 t} \cdot [B]\right) = k_1 \cdot A_0 \int_0^t e^{(k_2 - k_1)t} \mathrm{d}t \qquad \text{eq. 8.48}$$

which gives

$$e^{k_2 t} \cdot [B](t) = \frac{k_1 \cdot A_0}{k_2 - k_1}\left(e^{(k_2 - k_1)t} - 1\right) \qquad \text{eq. 8.49}$$

Multiplication of eq. 8.49 with exp(-k_2t) then yields the integrated rate law

$$[B](t) = A_0 \frac{k_1}{k_2 - k_1}\left(e^{-k_1 t} - e^{-k_2 t}\right) \qquad \text{eq. 8.50}$$

Finally, $[C](t)$ can be calculated from the mass conservation as $[C](t) = A_0 - [A](t) - [B](t)$ as

$$[C](t) = A_0\left(1 + \frac{k_1 e^{-k_2 t} - k_2 e^{-k_1 t}}{k_2 - k_1}\right) \qquad \text{eq. 8.51}$$

The time dependence of $[A]$, $[B]$, and $[C]$ is shown in Figure 8.4. B is only transiently produced as an intermediate. The degree of accumulation of B depends on the ratio of k_1 and k_2: if $k_1 > k_2$, the reaction from B to C is slower than the reaction from A to B, and B will accumulate. In this case, B is formed with the rate constant k_1, and then slowly disappears with rate constant k_2.

If $k_1 < k_2$, then the reaction from B to C is faster than the reaction from A to B. Thus, any B that has been formed will rapidly react to C, and B will not accumulate as an intermediate. In this case, B is formed with the rate constant k_2, and decomposes with k_1. This intrinsically non-intuitive behavior becomes clear if we look at the equation for $[B](t)$ (eq. 8.50), and its behavior for the two cases: the amplitude factor contains the difference $k_2 - k_1$ in the denominator. This difference will be positive for $k_1 < k_2$. Consequently, the first exponential term with k_1 will stay positive, and will describe an exponential decay. The second exponential term with k_2 will remain negative, describing exponential growth. Thus, B will be formed rapidly with k_2, and decay slowly with k_1. However, if $k_1 > k_2$, the sign of the denominator becomes negative, converting the first exponential term with k_1 into an exponential growth, and the second exponential term with k_2 into an exponential decay. Thus, in this case B will be formed rapidly with k_1, and decay slowly with k_2.

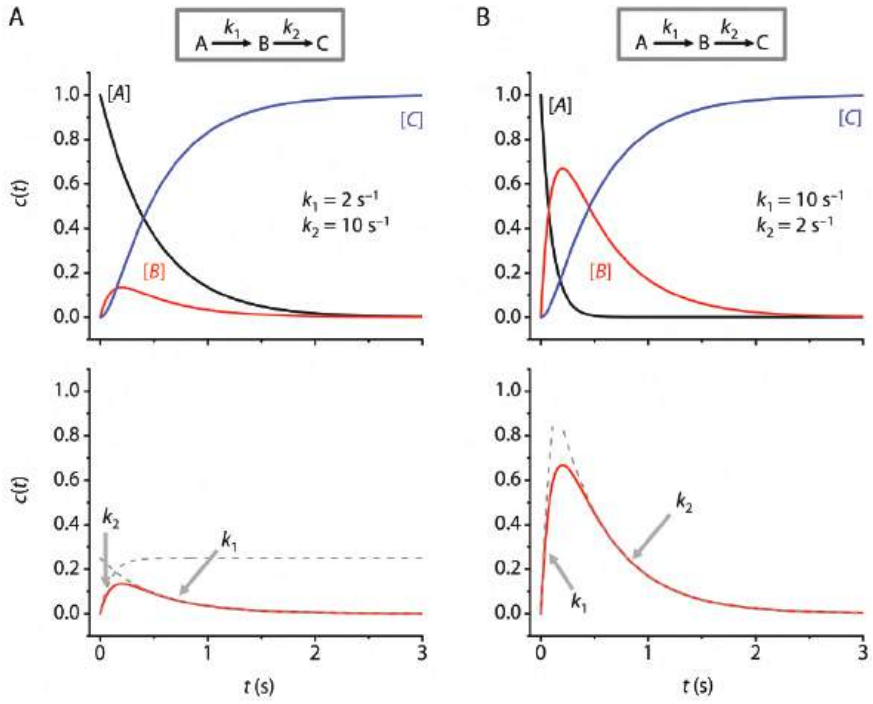


FIGURE 8.4: Consecutive reactions. A: Consecutive reactions from A to B to C with rate constants $k_1 < k_2$ ($k_1 = 2\ s^{-1}$, $k_2 = 10\ s^{-1}$). The concentration of A decreases exponentially with k_1. The concentration of B shows a maximum. The maximum concentration is reached with k_2, and decays with k_1 (bottom). The maximum concentration of B depends on A_0 and the ratio of $k_1/(k_1+k_2)$. C is formed after a short lag period after A has been converted to B. B: Consecutive reactions with rate constants $k_1 > k_2$ ($k_1 = 10\ s^{-1}$, $k_2 = 2\ s^{-1}$). Because the reaction consuming B is now slower than the reaction producing B, B accumulates transiently. The maximum concentration of B is reached with k_1, and decays with k_2 (bottom).

B_{max} can be calculated from the first derivative of eq. 8.50: at the maximum of the curve describing $[B](t)$ the first derivative will be zero (eq. 8.36)

$$\frac{d[B](t)}{dt} = k_1[A] - k_2[B] = 0 \qquad \text{eq. 8.36}$$

or

$$k_1[A] = k_2[B] \qquad \text{eq. 8.52}$$

We can rearrange this equation to obtain

$$k_1\left(A_0 - [B]_{max}\right) = k_2[B]_{max} \qquad \text{eq. 8.53}$$

and solve for $[B]_{max}$, the maximum concentration of B:

$$[B]_{max} = \frac{k_1}{k_1 + k_2} A_0 \qquad \text{eq. 8.54}$$

According to this equation, we can neglect k_2 in the denominator for $k_1 \gg k_2$, and $[B]_{max}$ will approach A_0 (Figure 8.4). For $k_1 \ll k_2$, we can neglect k_1 in the denominator. $[B]_{max}$ will then reach

$$[B]_{max} \approx \frac{k_1}{k_2} A_0 \qquad \text{eq. 8.55}$$

In this case, the ratio of the two individual rate constants determines the amount of B that accumulates (Figure 8.5). The faster the formation of B compared to its decomposition, the higher $[B]_{max}$

(Box 8.1). Using eq. 8.50, we can also determine the time t_{max} when the intermediate B reaches the maximal concentration. The time t_{max} is the ideal time to capture an unstable reaction intermediate, either to characterize it or to study its reaction with other compounds.

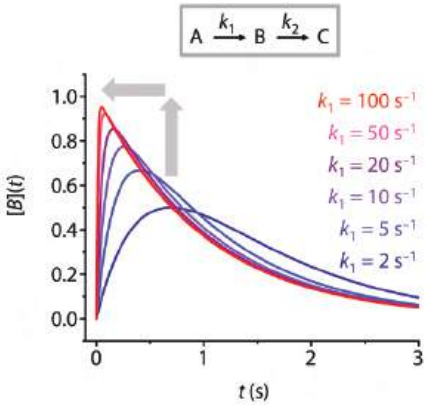


FIGURE 8.5: Accumulation of B as a function of k_1 and k_2. The concentration of B as a function of time is shown for k_1 = 2 s^{-1}, 5 s^{-1}, 10 s^{-1}, 20 s^{-1}, 50 s^{-1}, and 100 s^{-1}. For all curves, k_2 is 1 s^{-1}. The faster the first reaction, the more B is transiently accumulated, and the maximum concentration of B is reached earlier (gray arrows).

BOX 8.1: ACCUMULATION OF INTERMEDIATES IN SEQUENTIAL REACTIONS: ENZYMATIC OXIDATION OF ETHANOL.

Ethanol is metabolized by oxidation (Scheme 8.5). In a first step, alcohol dehydrogenase (ADH) oxidizes ethanol to acetaldehyde. Acetaldehyde is then oxidized to acetic acid by acetaldehyde dehydrogenase (ALDH).



scheme 8.5

Acetaldehyde is toxic and causes nausea, headaches, and tachycardia. Normally, the intermediate acetaldehyde does not accumulate because its oxidation is faster than its generation by ethanol oxidation. However, in people with mutations in ALDH, acetaldehyde oxidation is slower, and acetaldehyde accumulates to higher levels. ALDH dehydrogenase inhibitors are used in the treatment of alcoholism to increase the negative effects and discourage patients from drinking.

In principle, we have now treated the key elements of complex reaction pathways and networks. While the definition of reaction velocities for complex reaction networks follows the same basic principle as for the building blocks we have treated in Chapters 7 and 8, the derivation of integrated rate laws becomes more and more challenging. From the set of differential equations that describe the reaction network, the temporal behavior of the entire network can be modeled using numerical methods (Box 8.2).

BOX 8.2: DIFFERENTIAL EQUATIONS TO DESCRIBE REACTION NETWORKS.

In reaction networks, the same compound may be a reactant for one reaction, and a product of another reaction. Although this and the lack of directionality pose a problem in defining a reaction velocity, the temporal behavior of such a network can be described quantitatively. Even for the most complicated network, we can express the changes in concentration for each of the reactants in the form of differential equations. We only have to consider all processes in which this compound is involved, and separate them into processes toward the compound that cause an increase of its concentration, and all processes that lead away from this compound and cause a concentration decrease (Figure 8.6).

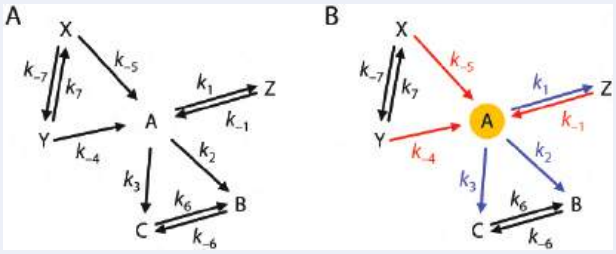


FIGURE 8.6: A reaction network. A: The compounds A, B, C, X, Y, and Z are connected through several reversible and irreversible reactions. B: To describe the change in concentration of A over time, we need to take into account all processes that lead away from A and cause a decrease in its concentration (blue, with a negative sign, and all processes that lead to formation of A and cause an increase in the concentration of A (red), with a positive sign.

We first focus on the balance for compound A (Figure 8.6). All processes that lead away from A contribute to a concentration decrease, and need to be taken into account with a negative sign in the expression for $d[A]/dt$. In contrast, all processes towards A increase the concentration of A and are taken into account with a positive sign:

$$\frac{d[A]}{dt} = -k_1[A] - k_2[A] - k_3[A] + k_{-1}[Z] + k_{-4}[Y] + k_{-5}[X] \qquad \text{eq. 8.56}$$

Similarly, we can formulate the change in the concentration of B as

$$\frac{d[B]}{dt} = -k_{-6}[B] + k_2[A] + k_6[C] \qquad \text{eq. 8.57}$$

and can express the changes in concentrations for all other compounds by differential equations. Such a set of differential equations can be solved analytically, for example by using matrix methods (Chapter 9). Alternatively, the temporal behavior of the reaction network can be calculated from these equations using numerical methods, provided that the individual rate constants are known. This approach is the core of many systems biology approaches to model metabolic or signaling networks. The predictive power for changes in the entire network in response to specific changes depends on how well determined the rate constants, the connections between individual compounds, and their starting concentrations are.

We will see in Chapter 14 how we can apply the principles we have learnt from model reactions to quantitatively understand more complex multi-step reactions.

QUESTIONS

8.1 A reversible reaction A + B ⇄ AB can be monitored by fluorescence because the fluorescence of B increases upon AB formation. Binding is too fast to be measured and already complete within the dead time of your experiment. How can you determine K_{eq}, k_{-1}, and k_1?

8.2 You start from a reactant A and monitor formation of the product B of a reaction as a function of time. $[B](t)$ can be described by a single exponential function with a rate constant $k = 0.4\ s^{-1}$. The end concentration of B is 7.5 μM. What are the possible explanations for this behavior and what is the meaning of the measured rate constant in each case? Which piece of information is needed to distinguish between the different possibilities?

8.3 A reactant A is converted to B with k_1, B is converted to C with k_2. A also reacts to D with k_3. What determines the ratio of C and D at the end of the reaction? What determines the time dependence $[B](t)$?

8.4 An enzyme binds and hydrolyzes ATP. You want to populate the enzyme-ATP complex to study its binding to an interaction partner. ATP binding induces a conformational change in the enzyme, and the overall binding reaction is slow with $k_{bind} = 0.1\ s^{-1}$. The rate constant of ATP hydrolysis is $k_{hyd} = 10^{-3}\ s^{-1}$. How long do you have to incubate the enzyme with ATP to maximize the concentration of the ATP-bound form?

8.5 For a FRET experiment, you want to label a protein with two cysteines with a donor and an acceptor fluorophore. The reactions of both cysteines are irreversible; cysteine 1 reacts with both dyes with $k_1 = 1\ s^{-1}$, cysteine 2 with $k_2 = 0.1\ s^{-1}$. For labeling, you incubate with the donor dye first, then stop the reaction at t_1. The reaction products are then labeled to completion with the acceptor dye. Calculate the fractions of DD, DA, AD- and AA-labeled species for a reaction time t_1 of one, seven, or ten half-lives of cysteine 1.

8.6 For the reaction scheme in Question 8.5, what is the incubation time with the donor dye that gives the maximum fraction of DA-species, the maximum ratio of DA/AD-species, or the maximum for the sum of the DA- and AD-labeled species?

8.7 An intermediate I is on the pathway from A to B (reaction 1). The same intermediate is generated on the reaction pathway from C to the product D (reaction 2). The conversion from A to I is 10-fold faster than the conversion from C to I, the conversion from I to B is 2-fold faster than the conversion from I to D. Which reaction accumulates more intermediate? What determines how much more B is accumulated? Which reaction accumulates I earlier in the reaction?

REFERENCES

Hald, J. and Jacobsen, E. (1948) A drug sensitizing the organism to ethyl alcohol. *Lancet* **2**(6539): 1001–1004.
· treatment of alcoholism with disulfiram/Antabuse

Thornburg, Z. R., Bianchi, D. M., Brier, T. A., Gilbert, B. R., Earnest, T. M., …, Luthey-Schulten, Z. (2022) Fundamental behaviors emerge from simulations of a living minimal cell. *Cell* **185**: 345–360.
· presents a whole-cell fully dynamic kinetic model of a minimal cell

9

Deriving integrated rate laws by solving sets of differential equations with matrix algebra

So far, we have derived integrated rate laws by writing down the complete set of differential equations for all species in the reaction scheme considered, followed by integration. An alternative approach uses matrix algebra to determine integrated rate laws. This approach is particularly helpful for more complex systems where the direct integration is difficult or cumbersome (see Box 8.2). In the following, the principle of this approach is illustrated for reversible reactions, for consecutive, irreversible reactions, and for consecutive reactions with a first reversible and a second, irreversible step.

9.1 REVERSIBLE REACTIONS

The scheme for a reversible reaction is

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B$$ scheme 8.1

The set of differential equations describing this system is

$$\frac{d[A]}{dt} = -k_1[A] + k_{-1}[B]$$ eq. 8.2

$$\frac{d[B]}{dt} = k_1[A] - k_{-1}[B]$$ eq. 9.1

This is a coupled system of differential equations, since determination of $[A](t)$ from eq. 8.2 requires knowledge of $[B](t)$, and solving eq. 9.1 for $[B](t)$ requires knowledge of $[A](t)$. In Section 8.1, the integrated rate laws for $[A](t)$ (eq. 8.17) and $[B](t)$ (eq. 8.18) are obtained by separating the variables, substitution, integration, and reversal of substitution, along with applying the initial conditions and the law of mass action. Here, we will now work our way to the same rate laws using the matrix approach.

The general solution of a linear differential equation with the constant λ

$$x(t)' = \lambda \cdot x(t) \qquad \text{eq. 9.2}$$

is

$$x(t) = c \cdot e^{\lambda t} + offset \qquad \text{eq. 9.3}$$

The offset or integration constant may assume any value, so there are infinitely many solutions to eq. 9.2. The actual value of the offset will be determined in a later step, when the starting conditions are applied. We may therefore disregard the offset until we reach that stage in our calculations. The factor c is also determined by the starting conditions. Both the offset and c are independent of time. To extend the expressions in eq. 9.2 and eq. 9.3 from a single equation to a set of n differential equations, we summarize the n coefficients λ in a $n \times n$ matrix M. In analogy to eq. 9.2, we can then write

$$\vec{x}(t)' = M \cdot \vec{x}(t) \qquad \text{eq. 9.4}$$

where $\vec{x}(t)$ is a vector whose components are the unknown concentrations of all species of the system as a function of time. In the notation of eq. 9.4, the system of equations for a reversible reaction described by eq. 8.2 and eq. 9.1 looks like this:

$$\vec{x}(t)' = \begin{pmatrix} \frac{\mathrm{d}[A]}{\mathrm{d}t} \\ \frac{\mathrm{d}[B]}{\mathrm{d}t} \end{pmatrix} = M \cdot \vec{x}(t) = \begin{pmatrix} -k_1 & k_{-1} \\ k_1 & -k_{-1} \end{pmatrix} \cdot \begin{pmatrix} [A] \\ [B] \end{pmatrix} \qquad \text{eq. 9.5}$$

The analogous equation to eq. 9.3 is

$$\vec{x}(t) = \vec{\eta} \cdot e^{\lambda t} \qquad \text{eq. 9.6}$$

where $\vec{\eta}$ is a vector stating the pre-exponential factors or amplitudes. λ and $\vec{\eta}$ are termed *Eigenvalues* and *Eigenvectors* of the matrix M. Section 31.10.3 describes how Eigenvalues are determined from the determinant of a matrix and used to obtain Eigenvectors. Eigenvalues and Eigenvectors provide a general recipe to solve problems such as the integrated rate law problem stated in eq. 9.4: we first have to calculate the Eigenvalues λ of the matrix M. We then use these to determine the Eigenvectors $\vec{\eta}$ of M. Finally, we put the Eigenvectors $\vec{\eta}$ into eq. 9.6 to arrive at the integrated rate laws.

To determine the Eigenvalues and Eigenvectors, we first need to bring our problem into a suitable form that contains the matrix M, its Eigenvalues λ, and its Eigenvectors $\vec{\eta}$. We achieve this by combining eq. 9.4 and eq. 9.6 to eliminate terms $x(t)$ and $x'(t)$. Differentiation of eq. 9.6 leads to a description for $x'(t)$ by exponential functions:

$$\vec{x}(t)' = \lambda \cdot \vec{\eta} \cdot e^{\lambda t} \qquad \text{eq. 9.7}$$

Placing the resulting eq. 9.7 into eq. 9.4 gives:

$$\lambda \cdot \vec{\eta} \cdot e^{\lambda t} = M \cdot \vec{x}(t) \qquad \text{eq. 9.8}$$

For $\vec{x}(t)$ we have an expression in eq. 9.6, so we can replace all functions $x(t)$ in eq. 9.8 with exponential functions:

$$\lambda \cdot \vec{\eta} \cdot e^{\lambda t} = M \cdot \vec{\eta} \cdot e^{\lambda t} \qquad \text{eq. 9.9}$$

Eq. 9.9 is now a function of M, λ, and $\vec{\eta}$. It can be rearranged to

$$\left(M \cdot \vec{\eta} - \lambda \cdot \vec{\eta}\right) \cdot e^{\lambda t} = \vec{0} \qquad \text{eq. 9.10}$$

Factoring out the vector $\vec{\eta}$ results in eq. 9.11:

$$\left(M - \lambda\right) \cdot \vec{\eta} \cdot e^{\lambda t} = \vec{0} \qquad \text{eq. 9.11}$$

The exponential term will never be zero, so eq. 9.11 simplifies to

$$\left(M - \lambda\right) \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.12}$$

The only problem left is that this equation would require us to subtract a constant λ from a matrix M, which is impossible: only matrices can be subtracted from one another. The subtraction becomes possible when expanding the scalar λ with the identity matrix I (Section 31.10.1). Matrix I is chosen to have the same dimensions (number of rows and columns) as matrix M. Its inclusion does nothing but multiplying all elements of the vector $\vec{\eta}$ by 1 and increasing the dimension of the term $\lambda \cdot \vec{\eta}$ to that of the first term, $M \cdot \vec{\eta}$, so that the subtraction can be done. Including I in eq. 9.12 yields

$$\left(M - \lambda \cdot I\right) \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.13}$$

Eq. 9.13 is the general condition for Eigenvalues and Eigenvectors. We could have gotten to this point more quickly without regarding the actual problem, but while keeping the view on the kinetic problem it now shows that the method is applicable universally, since all $e^{\lambda t}$ terms in eq. 9.11, which contain the kinetic information, can be disregarded.

To apply eq. 9.13 to our specific problem, we insert the matrix M from eq. 9.5 into eq. 9.13:

$$\left[\begin{pmatrix} -k_1 & k_{-1} \\ k_1 & -k_{-1} \end{pmatrix} - \lambda \cdot I\right] \cdot \vec{\eta} = \left[\begin{pmatrix} -k_1 & k_{-1} \\ k_1 & -k_{-1} \end{pmatrix} - \lambda \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.14}$$

The term in the brackets of eq. 9.14 simplifies to give eq. 9.15:

$$\begin{pmatrix} -k_1 - \lambda & k_{-1} \\ k_1 & -k_{-1} - \lambda \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.15}$$

We then use eq. 9.15 to obtain λ and $\vec{\eta}$ for the matrix M in eq. 9.6. The Eigenvalues λ of a matrix M are obtained by setting its determinant (Section 31.10.3) to zero, which for a 2×2 matrix gives a quadratic equation:

$$det\begin{pmatrix} -k_1 - \lambda & k_{-1} \\ k_1 & -k_{-1} - \lambda \end{pmatrix} = 0 = \left(-k_1 - \lambda\right)\left(-k_{-1} - \lambda\right) - k_1 k_{-1} = \lambda^2 + \left(k_1 + k_{-1}\right) \cdot \lambda \qquad \text{eq. 9.16}$$

The quadratic equation has the solutions $\lambda_1 = 0$ and $\lambda_2 = -k_1 - k_{-1}$, which are the two Eigenvalues we were looking for. In the next step, we now use these values to determine the two Eigenvectors, $\vec{\eta}_1$ and $\vec{\eta}_2$, one for each λ, by using eq. 9.13:

$$\left(M-\lambda\cdot I\right)\cdot\vec{\eta}=\begin{pmatrix}-k_1-\lambda & k_{-1}\\ k_1 & -k_{-1}-\lambda\end{pmatrix}\cdot\vec{\eta}=\vec{0} \qquad \text{eq. 9.17}$$

For $\lambda_1 = 0$ we determine $\vec{\eta}_1$ by solving the corresponding system of linear equations given by eq. 9.13:

$$\begin{pmatrix}-k_1 & k_{-1}\\ k_1 & -k_{-1}\end{pmatrix}\cdot\vec{\eta}=\begin{pmatrix}-k_1 & k_{-1}\\ k_1 & -k_{-1}\end{pmatrix}\cdot\begin{pmatrix}n_1\\ n_2\end{pmatrix}=\vec{0} \qquad \text{eq. 9.18}$$

The two equations are linearly dependent (*i.e.* one can be converted into the other by multiplication with –1):

$$\begin{aligned}-k_1\cdot n_1+k_{-1}\cdot n_2=0\\ k_1\cdot n_1-k_{-1}\cdot n_2=0\end{aligned} \qquad \text{eq. 9.19}$$

Thus, it is sufficient if we solve one of these equations. We obtain infinitely many solutions of the form:

$$n_1=\frac{k_{-1}}{k_1}\cdot n_2 \qquad \text{eq. 9.20}$$

By choosing $n_2 = 1$, we obtain one form of the Eigenvector $\vec{\eta}_1$ as

$$\vec{\eta}_1=c_1\cdot\begin{pmatrix}n_1\\ n_2\end{pmatrix}=c_1\cdot\begin{pmatrix}k_{-1}/k_1\\ 1\end{pmatrix} \qquad \text{eq. 9.21}$$

where c_1 is a variable factor, which we will need later to factor in the starting conditions $[A](0) = A_0$ and $B_0 = 0$ (the "offset" in eq. 9.3). Note that we could also have taken $n_2 = k_1$, in which case the Eigenvector would be written as

$$\vec{\eta}_1=c_1\cdot\begin{pmatrix}k_{-1}\\ k_1\end{pmatrix} \qquad \text{eq. 9.22}$$

with a different value for constant c_1. As mentioned above, the vectors in eq. 9.21 and eq. 9.22 are linearly dependent since they are only different by a constant (linear) factor:

$$\begin{pmatrix}k_{-1}\\ k_1\end{pmatrix}=k_1\cdot\begin{pmatrix}k_{-1}/k_1\\ 1\end{pmatrix} \qquad \text{eq. 9.23}$$

For the second Eigenvector $\vec{\eta}_2$, associated with Eigenvalue $\lambda_2 = -k_1-k_{-1}$, we again solve the system of equations given by eq. 9.13:

$$\begin{pmatrix}-k_1-\lambda & k_{-1}\\ k_1 & -k_{-1}-\lambda\end{pmatrix}\cdot\vec{\eta}=\begin{pmatrix}-k_1-\left(-k_1-k_{-1}\right) & k_{-1}\\ k_1 & -k_{-1}-\left(-k_1-k_{-1}\right)\end{pmatrix}\cdot\begin{pmatrix}n_1\\ n_2\end{pmatrix}=\vec{0} \qquad \text{eq. 9.24}$$

which reduces to

$$\begin{pmatrix}k_{-1} & k_{-1}\\ k_1 & k_1\end{pmatrix}\cdot\begin{pmatrix}n_1\\ n_2\end{pmatrix}=\begin{pmatrix}0\\ 0\end{pmatrix} \qquad \text{eq. 9.25}$$

Using the Gaussian method for row reduction of matrices (Section 31.10.2), we multiply the first row by k_1, and the second row by $-k_{-1}$, then sum the two rows to obtain a new second row:

$$\begin{pmatrix} k_{-1} & k_{-1} \\ 0 & k_{-1}k_1 - k_1k_{-1} \end{pmatrix} \cdot \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \begin{pmatrix} k_{-1} & k_{-1} \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} n_1 \\ n_2 \end{pmatrix} = \vec{0} \qquad \text{eq. 9.26}$$

Again, we obtain infinitely many solutions, this time of the form $n_1 = -n_2$, and by choosing $n_1 = 1$, the second Eigenvector is

$$\vec{\eta}_2 = c_2 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} \qquad \text{eq. 9.27}$$

The next step is using the Eigenvectors and Eigenvalues to obtain the solutions to the system of equations stated by eq. 9.6. We obtain two sets of solutions because we have two Eigenvalues and their corresponding Eigenvectors:

$$\vec{x}_1(t) = \vec{\eta}_1 \cdot e^{\lambda_1 t}$$

$$\vec{x}_2(t) = \vec{\eta}_2 \cdot e^{\lambda_2 t} \qquad \text{eq. 9.28}$$

These two solutions form a *fundamental set* of solutions: both satisfy eq. 9.6 and they are independent from one another, *i.e.* cannot be converted into each other by a simple factor. For a system of linear differential equations (eq. 9.4), which we are considering here, the *general solution* is then defined as the sum of these fundamental solutions:

$$\vec{x}(t) = \sum_i c_i \cdot \vec{\eta}_i \cdot e^{\lambda_i \cdot t} = \vec{\eta}_1 \cdot e^{\lambda_1 t} + \vec{\eta}_2 \cdot e^{\lambda_2 t} \qquad \text{eq. 9.29}$$

In eq. 9.29, $\vec{x}(t)$ represents the set of integrated rate laws we want to determine, the time dependence of the concentrations of species A and B, or $[A](t)$ and $[B](t)$, as a function of the rate constants k_1 and k_{-1}. By inserting the Eigenvalues and Eigenvectors, eq. 9.29 becomes

$$\vec{x}(t) = c_1 \cdot \begin{pmatrix} k_{-1}/k_1 \\ 1 \end{pmatrix} \cdot e^{0 \cdot t} + c_2 \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} \cdot e^{-(k_1 + k_{-1}) \cdot t} \qquad \text{eq. 9.30}$$

Written as individual equations for $[A](t)$ and $[B](t)$, eq. 9.30 becomes:

$$[A](t) = c_1 \cdot \frac{k_{-1}}{k_1} + c_2 \cdot e^{-(k_1 + k_{-1}) \cdot t} \qquad \text{eq. 9.31}$$

and

$$[B](t) = c_1 - c_2 \cdot e^{-(k_1 + k_{-1}) \cdot t} \qquad \text{eq. 9.32}$$

The last part is to determine the values for c_1 and c_2 from the starting conditions (the "offset" in eq. 9.3). From the boundary condition $[B](0) = 0$, we obtain from eq. 9.32:

$$0 = c_1 - c_2 \qquad \text{eq. 9.33}$$

or

$$c_1 = c_2 \qquad \text{eq. 9.34}$$

Similarly, inserting the boundary condition $[A](0) = A_0$ into eq. 9.31 leads to

$$A_0 = c_1 \cdot \frac{k_{-1}}{k_1} + c_2 = c_1 \cdot \frac{k_{-1} + k_1}{k_1} \qquad \text{eq. 9.35}$$

or, solved for c_1:

$$c_1 = A_0 \cdot \frac{k_1}{k_{-1} + k_1} \qquad \text{eq. 9.36}$$

Replacing c_1 and c_2 in eq. 9.31 and eq. 9.32 gives

$$[A](t) = A_0 \cdot \frac{k_1}{k_{-1} + k_1} \cdot \frac{k_{-1}}{k_1} + A_0 \cdot \frac{k_1}{k_{-1} + k_1} \cdot e^{-(k_1 + k_{-1}) \cdot t} \qquad \text{eq. 9.37}$$

and

$$[B](t) = \frac{A_0 \cdot k_1}{k_{-1} + k_1} \cdot \left(1 - e^{-(k_1 + k_{-1}) \cdot t}\right) \qquad \text{eq. 9.38}$$

Canceling k_1 and factoring out $A_0/(k_1+k_{-1})$ in eq. 9.37 results in

$$[A](t) = \frac{A_0}{k_{-1} + k_1} \cdot \left(k_{-1} + k_1 \cdot e^{-(k_1 + k_{-1}) \cdot t}\right) \qquad \text{eq. 9.39}$$

As expected, both integrated rate laws are identical to the rate laws obtained from direct integration of the individual differential equations (see Section 8.1, eq. 8.17 and eq. 8.18). Direct integration is quicker in simpler cases but often impossible in more difficult kinetic scenarios. The admittedly lengthy pathway using Eigenvalues and Eigenvectors has the advantage of being generally applicable.

9.2 CONSECUTIVE, IRREVERSIBLE REACTIONS

The reaction scheme for consecutive, irreversible reactions is

$$\mathrm{A} \xrightarrow{k_1} \mathrm{B} \xrightarrow{k_2} \mathrm{C} \qquad \text{scheme 8.4}$$

The differential equations that need to be solved are

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k_1[A] \qquad \text{eq. 8.35}$$

$$\frac{\mathrm{d}[B]}{\mathrm{d}t} = k_1[A] - k_2[B] \qquad \text{eq. 8.36}$$

and

$$\frac{\mathrm{d}[C]}{\mathrm{d}t} = k_2[B] \qquad \text{eq. 8.37}$$

In Section 8.3, we solved this coupled system of three differential equations in two unknowns by first integrating eq. 8.35, and then substituting the result into eq. 8.36, followed by integration. The solution to eq. 8.37 was obtained by applying the law of mass conservation (see Section 8.3).

To solve the system of differential equations with the matrix approach, we transfer eq. 9.37–eq. 9.39 into the matrix representation:

$$\vec{x}' = \begin{pmatrix} \frac{\mathrm{d}[A]}{\mathrm{d}t} \\ \frac{\mathrm{d}[B]}{\mathrm{d}t} \\ \frac{\mathrm{d}[C]}{\mathrm{d}t} \end{pmatrix} = M \cdot \vec{x} = \begin{pmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix} \cdot \begin{pmatrix} [A] \\ [B] \\ [C] \end{pmatrix} \qquad \text{eq. 9.40}$$

We then determine the Eigenvalues of the 3×3 matrix M, whose determinant leads to the characteristic cubic polynomial:

$$det \begin{pmatrix} -k_1 - \lambda & 0 & 0 \\ k_1 & -k_2 - \lambda & 0 \\ 0 & k_2 & 0 - \lambda \end{pmatrix} = 0 = (-k_1 - \lambda)(-k_2 - \lambda)(-\lambda) \qquad \text{eq. 9.41}$$

From eq. 9.41, we obtain the Eigenvalues $\lambda_1 = 0$, $\lambda_2 = -k_2$ and $\lambda_3 = -k_1$. To determine the three Eigenvectors, we solve eq. 9.42, the analogous expression to eq. 9.17, for each λ:

$$(M - \lambda \cdot I) \cdot \vec{\eta} = \begin{pmatrix} -k_1 - \lambda & 0 & 0 \\ k_1 & -k_2 - \lambda & 0 \\ 0 & k_2 & 0 - \lambda \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.42}$$

For $\lambda_1 = 0$:

$$\begin{pmatrix} -k_1 & 0 & 0 \\ k_1 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.43}$$

which gives

$$\vec{\eta}_1 = c_1 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad \text{eq. 9.44}$$

For $\lambda_2 = -k_2$:

$$(M - \lambda \cdot I) \cdot \vec{\eta} = \begin{pmatrix} -k_1 + k_2 & 0 & 0 \\ k_1 & 0 & 0 \\ 0 & k_2 & k_2 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.45}$$

which gives

$$\vec{\eta}_2 = c_2 \cdot \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \qquad \text{eq. 9.46}$$

and for $\lambda_3 = -k_1$:

$$\left(M-\lambda\cdot I\right)\cdot\vec{\eta}=\begin{pmatrix}0 & 0 & 0\\ k_1 & -k_2+k_1 & 0\\ 0 & k_2 & k_1\end{pmatrix}\cdot\vec{\eta}=\vec{0} \qquad \text{eq. 9.47}$$

which gives

$$\vec{\eta}_3=c_3\cdot\begin{pmatrix}\left(k_1-k_2\right)/k_2\\ -k_1/k_2\\ 1\end{pmatrix} \qquad \text{eq. 9.48}$$

We now use the Eigenvalues and Eigenvectors to formulate the solution of eq. 9.40, which leads us to the integrated rate laws. The general solution is again the sum of the set of fundamental solutions:

$$\vec{x}_i\left(t\right)=\sum_i c_i\cdot\vec{\eta}_i\cdot e^{\lambda_i\cdot t}=c_1\cdot\vec{\eta}_1\cdot e^{\lambda_1\cdot t}+c_2\cdot\vec{\eta}_2\cdot e^{\lambda_2\cdot t}+c_3\cdot\vec{\eta}_3\cdot e^{\lambda_3\cdot t} \qquad \text{eq. 9.49}$$

With the Eigenvectors and Eigenvalues, we obtain the solution set

$$\vec{x}\left(t\right)=c_1\cdot\begin{pmatrix}0\\0\\1\end{pmatrix}\cdot e^{0\cdot t}+c_2\cdot\begin{pmatrix}0\\-1\\1\end{pmatrix}\cdot e^{-k_2\cdot t}+c_3\cdot\begin{pmatrix}\left(k_1-k_2\right)/k_2\\ -k_1/k_2\\ 1\end{pmatrix}\cdot e^{-k_1\cdot t} \qquad \text{eq. 9.50}$$

and in the expanded form

$$\left[A\right]\left(t\right)=c_1\cdot1\cdot0+c_2\cdot e^{-k_2\cdot t}\cdot0+c_3\cdot e^{-k_1\cdot t}\cdot\frac{k_1-k_2}{k_2}=c_3\cdot\frac{k_1-k_2}{k_2}\cdot e^{-k_1\cdot t} \qquad \text{eq. 9.51}$$

$$\left[B\right]\left(t\right)=c_1\cdot1\cdot0+c_2\cdot e^{-k_2\cdot t}\cdot\left(-1\right)+c_3\cdot e^{-k_1\cdot t}\cdot\left(-\frac{k_1}{k_2}\right)=-c_2\cdot e^{-k_2\cdot t}-c_3\cdot\frac{k_1}{k_2}\cdot e^{-k_1\cdot t} \qquad \text{eq. 9.52}$$

and

$$\left[C\right]\left(t\right)=c_1\cdot1\cdot1+c_2\cdot e^{-k_2\cdot t}\cdot1+c_3\cdot e^{-k_1\cdot t}\cdot1=c_1+c_2\cdot e^{-k_2\cdot t}+c_3\cdot e^{-k_1\cdot t} \qquad \text{eq. 9.53}$$

In the last step, we determine c_1, c_2, and c_3 from the boundary conditions, $[A](0) = A_0$, $B_0 = C_0 = 0$, which can also be written in matrix form as

$$\begin{pmatrix}0 & 0 & \frac{k_1-k_2}{k_2}\\ 0 & -1 & -\frac{k_1}{k_2}\\ 1 & 1 & 1\end{pmatrix}\cdot\begin{pmatrix}c_1\\c_2\\c_3\end{pmatrix}=\begin{pmatrix}A_0\\0\\0\end{pmatrix} \qquad \text{eq. 9.54}$$

We obtain c_3 from the first row in the matrix as $c_3=\frac{A_0\cdot k_2}{k_1-k_2}$. From eq. 9.52, we have $c_2=-\frac{k_1}{k_2}\cdot c_3$; substituting c_3 into this expression gives $c_2=-\frac{A_0\cdot k_1}{k_1-k_2}$. From eq. 9.53, we have the condition $c_1 = -c_2-c_3 = A_0$. Substituting these constants into eq. 9.51, eq. 9.52, and eq. 9.53 yields the three integrated rate laws:

$$\left[A\right]\left(t\right)=A_0\cdot e^{-k_1\cdot t} \qquad \text{eq. 9.55}$$

$$[B](t) = \frac{A_0 \cdot k_1}{k_1 - k_2} \cdot \left(e^{-k_2 \cdot t} - e^{-k_1 \cdot t}\right) \qquad \text{eq. 9.56}$$

and

$$[C](t) = A_0 \cdot \left(1 - \frac{k_1 \cdot e^{-k_2 \cdot t} - k_2 \cdot e^{-k_1 \cdot t}}{k_1 - k_2}\right) \qquad \text{eq. 9.57}$$

These are the same equations as eq. 8.38, eq. 8.50, and eq. 8.51 obtained by direct integration.

9.3 REVERSIBLE REACTION FOLLOWED BY AN IRREVERSIBLE SECOND STEP

The reaction scheme

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B \xrightarrow{k_2} C \qquad \text{scheme 9.1}$$

is characterized by the following set of linear differential equations:

$$\frac{d[A]}{dt} = -k_1[A] + k_{-1}[B] \qquad \text{eq. 9.58}$$

$$\frac{d[B]}{dt} = k_1[A] - (k_{-1} + k_2)[B] \qquad \text{eq. 9.59}$$

$$\frac{d[C]}{dt} = k_2[B] \qquad \text{eq. 9.60}$$

To solve these equations, we need the Eigenvalues and Eigenvectors for the 3×3 matrix M

$$M = \begin{pmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -(k_{-1} + k_2) & 0 \\ 0 & k_2 & 0 \end{pmatrix} \qquad \text{eq. 9.61}$$

The Eigenvalues are obtained from

$$det\begin{pmatrix} -k_1 - \lambda & k_{-1} & 0 \\ k_1 & -(k_{-1} + k_2) - \lambda & 0 \\ 0 & k_2 & -\lambda \end{pmatrix} = 0 \qquad \text{eq. 9.62}$$

giving the characteristic third-order polynomial

$$(-k_1 - \lambda) \cdot (-k_{-1} - k_2 - \lambda) \cdot (-\lambda) - k_1 k_{-1}(-\lambda) = 0 \qquad \text{eq. 9.63}$$

From eq. 9.63, we obtain the first Eigenvalue as $\lambda_1 = 0$, leaving a quadratic equation for the other two Eigenvalues ($\lambda \neq 0$):

$$\lambda^2 + (k_1 + k_{-1} + k_2) \cdot \lambda + k_1 k_{-1} = 0 = \lambda^2 + p \cdot \lambda + k_1 k_{-1} \qquad \text{eq. 9.64}$$

where we have substituted

$$p = k_1 + k_{-1} + k_2 \qquad \text{eq. 9.65}$$

The solutions for the quadratic equation are

$$\lambda_{2/3} = \frac{-p \pm \sqrt{p^2 - 4 \cdot k_1 k_{-1}}}{2} \qquad \text{eq. 9.66}$$

With

$$q = \sqrt{p^2 - 4 \cdot k_1 k_{-1}} \qquad \text{eq. 9.67}$$

we obtain the remaining two Eigenvalues as

$$\lambda_{2/3} = \frac{-p \pm q}{2} \qquad \text{eq. 9.68}$$

To obtain the Eigenvectors, we solve the matrix $(M–\lambda_i \cdot I)$ for each Eigenvalue:

$$\begin{pmatrix} -k_1 - \lambda & k_{-1} & 0 \\ k_1 & -(k_{-1} + k_2) - \lambda & 0 \\ 0 & k_2 & -\lambda \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.69}$$

For $\lambda_1 = 0$, eq. 9.69 re-writes to

$$\begin{pmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -(k_{-1} + k_2) & 0 \\ 0 & k_2 & 0 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.70}$$

We sum the first two rows of the matrix to obtain a new second line:

$$\begin{pmatrix} -k_1 & k_{-1} & 0 \\ 0 & -k_2 & 0 \\ 0 & k_2 & 0 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.71}$$

Summing the second and third rows yields:

$$\begin{pmatrix} -k_1 & k_{-1} & 0 \\ 0 & -k_2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.72}$$

The third row means that n_3 can assume any value, and we choose $n_3 = 1$. The second row requires that $-k_2 \cdot n_2 = 0$, so $n_2 = 0$. Likewise, the first row requires $n_1 = 0$. Therefore, the Eigenvector associated with $\lambda_1 = 0$ is:

$$\vec{\eta}_1 = c_1 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \qquad \text{eq. 9.73}$$

For $\lambda_2 = 0.5 \cdot (-p-q)$, the matrix is

$$\begin{pmatrix} -k_1 - \frac{-p-q}{2} & k_{-1} & 0 \\ k_1 & -(k_{-1}+k_2) - \frac{-p-q}{2} & 0 \\ 0 & k_2 & -\frac{-p-q}{2} \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.74}$$

We multiply the first row by k_1 and the second row by $(k_1+0.5\cdot(-p-q))$ and sum the two to obtain a new second row:

$$\begin{pmatrix} -k_1 - \frac{-p-q}{2} & k_{-1} & 0 \\ 0 & \left(-k_{-1} - k_2 - \frac{-p-q}{2}\right)\cdot\left(k_1 + \frac{-p-q}{2}\right) + k_1 k_{-1} & 0 \\ 0 & k_2 & -\frac{-p-q}{2} \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.75}$$

Using the notations for p and q, it turns out that the central term in row two is zero, which leaves

$$\begin{pmatrix} -k_1 + \frac{p+q}{2} & k_{-1} & 0 \\ 0 & 0 & 0 \\ 0 & k_2 & \frac{p+q}{2} \end{pmatrix} \cdot \vec{\eta} = \vec{0} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.76}$$

We can shift the second line down, and are done with Gaussian elimination:

$$\begin{pmatrix} -k_1 + \frac{p+q}{2} & k_{-1} & 0 \\ 0 & k_2 & \frac{p+q}{2} \\ 0 & 0 & 0 \end{pmatrix} \cdot \vec{\eta} = \vec{0} \qquad \text{eq. 9.77}$$

Again, the third row means that n_3 can assume any value, and we choose $n_3 = 1$. The second row requires

$$k_2 \cdot n_2 + \frac{p+q}{2} \cdot n_3 = 0 \qquad \text{eq. 9.78}$$

or with $n_3 = 1$

$$n_2 = \frac{-p-q}{2k_2} \qquad \text{eq. 9.79}$$

In order to obtain n_1, we use the condition in the second row of eq. 9.74:

$$k_1 \cdot n_1 - \left(k_{-1} + k_2 + \frac{-p-q}{2}\right) \cdot n_2 = 0 \qquad \text{eq. 9.80}$$

or

$$k_1 \cdot n_1 = \left(k_{-1} + k_2 + \frac{-p-q}{2}\right) \cdot n_2 \qquad \text{eq. 9.81}$$

which gives

$$n_1 = \frac{1}{k_1} \cdot \left(k_{-1} + k_2 + \frac{-p-q}{2} \right) \cdot \left(\frac{-p-q}{2k_2} \right) \qquad \text{eq. 9.82}$$

Using the notations for p and q, the right-hand side of eq. 9.82 simplifies to

$$n_1 = \left(\frac{p+q-2k_2}{2k_2} \right) \qquad \text{eq. 9.83}$$

yielding the Eigenvector

$$\vec{\eta}_2 = c_2 \cdot \begin{pmatrix} \frac{p+q-2k_2}{2k_2} \\ \frac{-p-q}{2k_2} \\ 1 \end{pmatrix} \qquad \text{eq. 9.84}$$

An analogous series of calculations for $\lambda_3 = 0.5 \cdot (-p+q)$ yields the third Eigenvector:

$$\vec{\eta}_3 = c_3 \cdot \begin{pmatrix} \frac{p-q-2k_2}{2k_2} \\ \frac{-p+q}{2k_2} \\ 1 \end{pmatrix} \qquad \text{eq. 9.85}$$

The general solution for the rate equations is then

$$\vec{x}(t) = c_1 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot e^{0 \cdot t} + c_2 \cdot \begin{pmatrix} \frac{p+q-2k_2}{2k_2} \\ \frac{-p-q}{2k_2} \\ 1 \end{pmatrix} \cdot e^{\frac{-p-q}{2} \cdot t} + c_3 \cdot \begin{pmatrix} \frac{p-q-2k_2}{2k_2} \\ \frac{-p+q}{2k_2} \\ 1 \end{pmatrix} \cdot e^{\frac{-p+q}{2} \cdot t} \qquad \text{eq. 9.86}$$

The starting conditions are $[A](0) = A_0$, $B_0 = C_0 = 0$, which can also be written in matrix form as

$$\begin{pmatrix} 0 & \frac{p+q-2k_2}{2k_2} & \frac{p-q-2k_2}{2k_2} \\ 0 & \frac{-p-q}{2k_2} & \frac{-p+q}{2k_2} \\ 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} A_0 \\ 0 \\ 0 \end{pmatrix} \qquad \text{eq. 9.87}$$

We move the last row to the top and sum the other two rows to obtain a new third row:

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & \frac{-p-q}{2k_2} & \frac{-p+q}{2k_2} \\ 0 & -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ A_0 \end{pmatrix} \qquad \text{eq. 9.88}$$

Multiplication of the last row with $(-p-q)/(2k_2)$ and summing with the second row yields

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & \frac{-p-q}{2k_2} & \frac{-p+q}{2k_2} \\ 0 & 0 & \frac{q}{k_2} \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ A_0 \cdot \frac{-p-q}{2k_2} \end{pmatrix} \qquad \text{eq. 9.89}$$

The last row gives c_3 as

$$c_3 = -\frac{A_0}{2} \cdot \left(\frac{p}{q} + 1 \right) \qquad \text{eq. 9.90}$$

From the last row in eq. 9.88, we obtain $c_2 = -A_0 - c_3$:

$$c_2 = -A_0 + \frac{A_0}{2} \cdot \left(\frac{p}{q} + 1 \right) = \frac{A_0}{2} \cdot \left(\frac{p}{q} - 1 \right) \qquad \text{eq. 9.91}$$

From the first row in eq. 9.88, *i.e.* $c_1 + c_2 + c_3 = 0$, we obtain

$$c_1 = -c_2 - c_3 = -\frac{A_0}{2} \cdot \left(\frac{p}{q} - 1 \right) + \frac{A_0}{2} \cdot \left(\frac{p}{q} + 1 \right) = A_0 \qquad \text{eq. 9.92}$$

The final integrated rate laws are then:

$$\vec{x}(t) = A_0 \cdot \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \cdot e^{0 \cdot t} + \frac{A_0}{2} \cdot \left(\frac{p}{q} - 1 \right) \cdot \begin{pmatrix} \frac{p+q-2k_2}{2k_2} \\ \frac{-p-q}{2k_2} \\ 1 \end{pmatrix} \cdot e^{\frac{-p-q}{2} \cdot t} - \frac{A_0}{2} \cdot \left(\frac{p}{q} + 1 \right) \cdot \begin{pmatrix} \frac{p-q-2k_2}{2k_2} \\ \frac{-p+q}{2k_2} \\ 1 \end{pmatrix} \cdot e^{\frac{-p+q}{2} \cdot t} \qquad \text{eq. 9.93}$$

or, in expanded and simplified form:

$$[A](t) = \frac{A_0 \cdot (q - p + 2k_1)}{2q} \cdot e^{\frac{-p-q}{2} \cdot t} - \frac{A_0 \cdot (-q - p + 2k_1)}{2q} \cdot e^{\frac{-p+q}{2} \cdot t} \qquad \text{eq. 9.94}$$

$$[B](t) = -\frac{A_0 k_1}{q} \cdot e^{\frac{-p-q}{2} \cdot t} + \frac{A_0 k_1}{q} \cdot e^{\frac{-p+q}{2} \cdot t} \qquad \text{eq. 9.95}$$

and

$$[C](t) = A_0 + \frac{A_0}{2} \cdot \left(\frac{p}{q} - 1 \right) \cdot e^{\frac{-p-q}{2} \cdot t} - \frac{A_0}{2} \cdot \left(\frac{p}{q} + 1 \right) \cdot e^{\frac{-p+q}{2} \cdot t} \qquad \text{eq. 9.96}$$

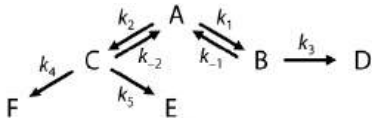
Now, we have to re-substitute p and q by the expressions defined above (eq. 9.65, eq. 9.67) to obtain the integrated rate laws describing the concentration of all components of the system as a function of time, with the initial concentration A_0 and the rate constants k_1, k_{-1}, and k_2 as parameters.

The matrix approach, illustrated in this Chapter for still rather simple reactions, can be applied to much larger systems of differential equations, describing increasingly complex reaction networks (see Box 8.2). The dimensionality of the matrix increases with the number of differential equations describing the system. While the manual diagonalization of the resulting n-dimensional matrix and the determination of the Eigenvalues and Eigenvectors becomes cumbersome with increasing n,

several software applications exist to solve large sets of such ordinary differential equations (ODEs). An alternative to the explicit solution of sets of differential equations is numerical integration. Both approaches are the key for systems biology, where they enable the quantitative, time-dependent modeling of the behavior of large cellular reaction networks, comprising parts of or even the entire cellular metabolism.

QUESTIONS

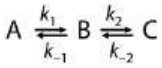
9.1 Formulate the differential equations and the matrix representation for the following reaction scheme:



9.2 Show with matrix algebra that an unfolded protein folds into its native structure with an apparent rate constant that is the sum of the rate constants for folding and unfolding.

9.3 Use matrix algebra to show that a reactant A that reacts in irreversible parallel reactions to the products B, C, or D with k_1, k_2, and k_3 decays with a rate constant that equals $k_{app} = k_1+k_2+k_3$. Use an online program such as Wolframalpha to obtain the Eigenvalues and Eigenvectors.

9.4 Derive the integrated rate laws for A, B, and C, linked by two consecutive, reversible reactions, using matrix algebra:



REFERENCES

Ikai, A. and Tanford, C. (1973) Kinetics of unfolding and refolding of proteins. I. Mathematical analysis. *J Mol Biol* **73**: 145–163.

· analytical solutions to a number of other equilibria

10

Rate-limiting steps

For a consecutive reaction from A *via* B to C according to Scheme 8.4

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \qquad \text{scheme 8.4}$$

we have derived in Section 8.3 that the formation of C follows eq. 8.51:

$$[C](t) = A_0\left(1 + \frac{k_1 e^{-k_2 t} - k_2 e^{-k_1 t}}{k_2 - k_1}\right) \qquad \text{eq. 8.51}$$

We now consider the limiting case that $k_1 \ll k_2$, *i.e.* the reaction from A to B is much slower than the reaction from B to C. With this limiting condition, the two exponential terms in eq. 8.51 will contribute very differently, with

$$k_1 e^{-k_2 t} << k_2 e^{-k_1 t} \qquad \text{eq. 10.1}$$

As an approximation, we can therefore ignore the first exponential term in the numerator. Similarly, we can simplify the denominator as

$$k_2 - k_1 \approx k_2 \qquad \text{eq. 10.2}$$

Altogether, eq. 8.51 then simplifies to

$$[C](t) \approx A_0\left(1 - \frac{k_2 e^{-k_1 t}}{k_2}\right) \qquad \text{eq. 10.3}$$

or

$$[C](t) \approx A_0\left(1 - e^{-k_1 t}\right) \qquad \text{eq. 10.4}$$

Thus, for the limiting condition that the first reaction is much slower than the second reaction in this sequence, C is formed approximately by an exponential function with the rate constant k_1 (Figure 10.1). The slow first step in the reaction sequence thus determines the overall reaction velocity from A to C. We therefore call the reaction from A to B with the rate constant k_1 the *rate-limiting step* of this sequence of reactions.

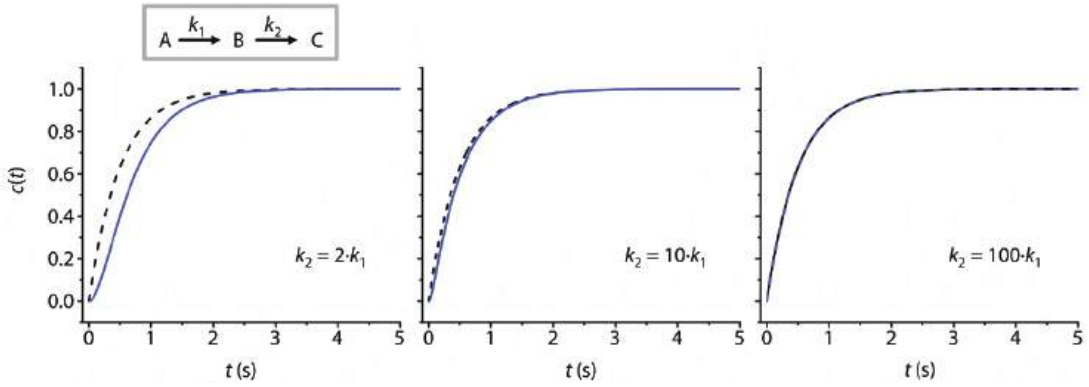


FIGURE 10.1: Rate-limiting steps. The slowest step in a consecutive reaction determines the rate of product formation. For two consecutive reactions A→B→C with rate constants k_1 and k_2, the formation of C can be described by an exponential increase with k_1 if $k_1 < k_2$. The dashed lines (black) represent the approximated single-exponential growth (eq. 10.4), the solid line (blue) shows the concentration of C calculated according to eq. 8.51 without approximation. The larger k_2 is compared to k_1, the more accurate is the description of $[C](t)$ by the approximation. $k_1 = 2\ s^{-1}$.

The rate-limiting step of a reaction sequence is often called the "bottleneck" of the pathway. The existence of a rate-limiting step in a reaction sequence has important implications for the response of the entire pathway to changes in reactant concentrations or to changes in the rates of individual steps. If we increase the rate constant k_2 in Scheme 8.4, *e.g.* by adding a catalyst for this reaction step (see Chapter 16), the overall reaction velocity will not be affected. If we, however, managed to accelerate the first reaction that constitutes the rate-limiting step, we would achieve an increase in the overall reaction velocity, resulting in a faster production of C. Overall reaction velocities are most efficiently altered by changing the velocity of the rate-limiting step(s) of reaction sequences! Hence, rate-limiting steps in biochemical reaction pathways are often subject to tight regulation (Box 10.1).

BOX 10.1: RATE-LIMITING STEPS IN METABOLIC PATHWAYS AND THEIR REGULATION.

Metabolic pathways can be regulated efficiently by accelerating rate-limiting steps. In a four-step reaction sequence A→B→C→D→E, the slowest step determines the rate of product generation ("bottleneck"). By up-regulating the enzyme that catalyzes the rate-limiting step, the rate of product formation and the overall flux through the pathway are increased (Figure 10.2). Up-regulation of any of the enzymes that catalyze faster steps has no effect on product formation rate and flux. If more than one of the reactions in the pathway are slow, acceleration of the slowest step only leads to an up-regulation until the second-slowest step becomes rate-limiting for the entire pathway. Efficient regulation of pathways is therefore only achieved by regulating the activities of several enzymes. In branched pathways, up-regulation of individual steps in one branch affects the kinetic partitioning between the two branches (Figure 10.2). Thus, regulation of individual steps in metabolic networks will have complex effects on different pathways and the balance of the entire network.

(*Continued*)

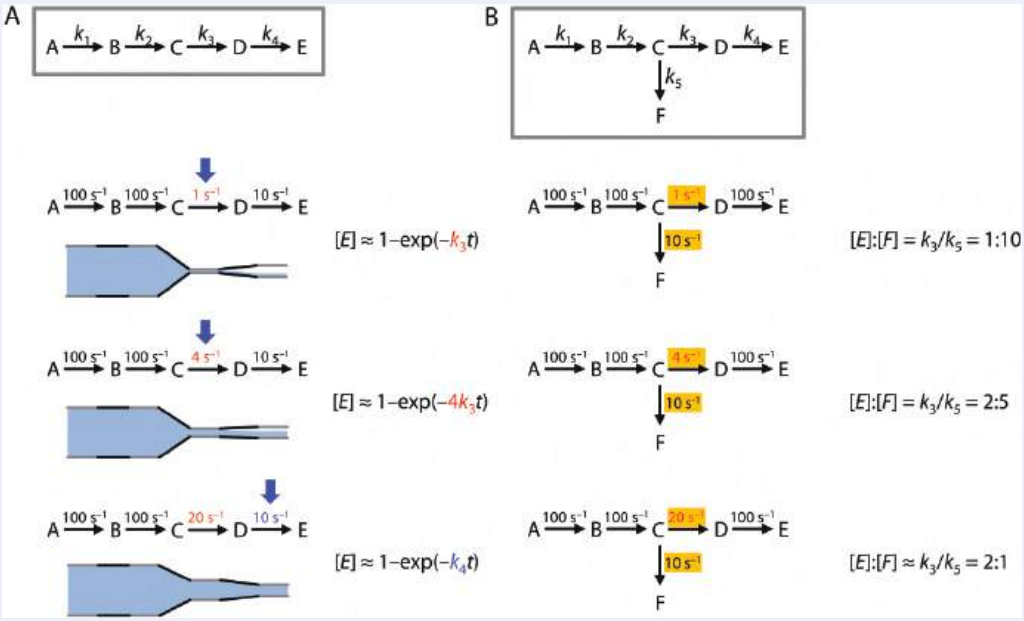


FIGURE 10.2: Rate-limiting steps and flux in metabolic pathways. A: The slowest reaction in a pathway (here: k_3) determines the rate of product formation and the flux (light blue) through the pathway. Acceleration of the rate-limiting step from $k_3 = 1\ s^{-1}$ to $k_3 = 4\ s^{-1}$ increases the rate of product formation and the overall flux. Further acceleration to $k_3 = 20\ s^{-1}$ leads to a further increase in the rate of product formation and to a change in the rate-limiting step (blue arrow). The reaction D→E is now rate-limiting: k_4 limits the rate of product formation. Further up-regulation of the enzyme that catalyzes the conversion of C to D does not lead to an increase in the rate of product formation or the overall flux. B: Same reaction sequence with a second branch from C to F. Regulation of the step C→D in the pathway from A to E affects the kinetic partitioning of C into the two alternative pathways, and the ratio of their end products E and F. The relevant rate constants for partitioning are highlighted in orange.

The identification of rate-limiting steps in a reaction sequence is important for the dissection of the reaction mechanism, and to understand the behavior of the overall system to changes in concentrations. Rate-limiting steps in biochemical reactions can be identified by the *kinetic isotope effect*. For example, a C-H bond in a substrate can be changed into a C-D bond (D = deuterium, ^2H). Due to the larger atomic mass of deuterium, the vibration frequency of the C-D bond is smaller than that of the C-H bond (see Section 21.4.1), and the zero-point energy of the C-D bond is lower than for the C-H bond. This decrease in energy will lead to an increase in activation energy for C-D bond breakage compared to breakage of a C-H bond. The higher activation energy leads to a smaller rate constant for bond breaking (Section 15.2): a C-D bond breaks more slowly than a C-H bond. The exchange of H to D will therefore slow down the overall reaction if breakage of this bond constitutes the rate-limiting step of the reaction (Figure 10.3), which is a *primary kinetic isotope effect*. If breakage of this bond is not the rate-limiting step, the overall reaction velocity will not be affected by the isotopic substitution.

Sometimes a kinetic isotope effect is measured even if the bond where the H-D substitution has been introduced is not the bond that breaks in the rate-limiting step. This effect is called *secondary kinetic isotope effect* (Box 10.2). A secondary isotope effect occurs when the substituted bond and the bond that breaks in the rate-limiting step are vibrationally coupled due to hyperconjugation.

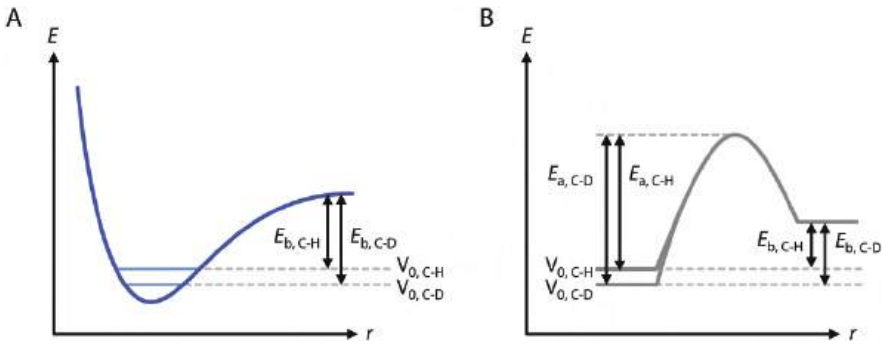


FIGURE 10.3: The kinetic isotope effect. A: Lennard-Jones potential (see Section 17.2.2) for a C-H and a C-D bond. The vibrational energy of the C-D bond is lower than for the C-H bond because of the larger mass of deuterium compared to hydrogen. B: Because of the lower energy (higher stability) of the C-D bond, the energy difference to the transition state (activation energy E_a, see Section 15.2) is higher for the C-D bond than for the C-H bond, and the rate constant for bond breaking is smaller.

The H/D kinetic isotope effect is large because an exchange of H for D doubles the mass, and has a large effect on bond vibration. Kinetic isotope effects based on substitution of ^{12}C by ^{13}C, for example, are much smaller due to the smaller relative change in mass.

BOX 10.2: PRIMARY AND SECONDARY KINETIC ISOTOPE EFFECTS AND ENZYME MECHANISMS.

The degradation of the non-polar amino acids methionine, isoleucine, and valine and of fatty acids with odd numbers of carbon atoms leads to propionyl-CoA. Propionyl-CoA is carboxylated to methylmalonyl-CoA, which then isomerizes to succinyl-CoA. The isomerization reaction is catalyzed by methylmalonyl-CoA mutase (Figure 10.4). Methylmalonyl-CoA mutase is a cobalamin-dependent enzyme. The cofactor 5'-deoxyadenosyl cobalamin (coenzyme B_{12}) is transiently converted into a radical during catalysis, and the isomerization follows a radical mechanism, which is very unusual in biochemistry. When the hydrogen in the methyl group of methylmalonyl-CoA is substituted by deuterium, a strong kinetic isotope effect (k_H/k_D = 3.5) is observed, pointing to abstraction of a hydrogen radical from this methyl group as the rate-limiting step of the catalyzed reaction. Miller & Richards (1969) J. Am. Chem. Soc. 91(6): 1498–1507.

Fumarase catalyzes the dehydration of malate to fumarate, a reaction at the end of the citric acid cycle, when oxaloacetate is regenerated from succinate. Fumarase catalyzes the stereospecific *trans*-addition of water to the C-C double bond (Figure 10.4). The mechanism has been analyzed using malate with a tritium at C2 or C3 as a substrate, and monitoring the enzyme-catalyzed dehydration to fumarate. For both substrates, a secondary kinetic isotope effect ($k_H/k_T \approx 1.14$) was observed, which is consistent with catalysis through a carbenium ion intermediate. Secondary kinetic isotope effects are generally used to provide evidence for carbenium ion intermediates. Schmidt *et al.* (1969) J. Am. Chem. Soc 91(21): 5849–5854.

(*Continued*)

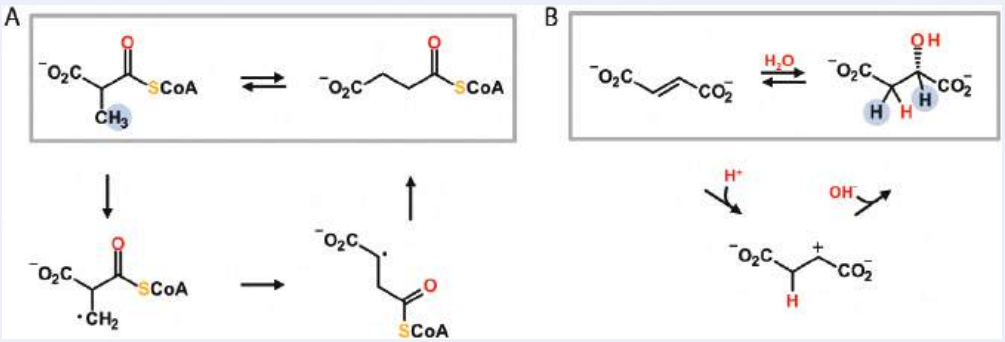


FIGURE 10.4: Isotope effects and mechanisms of enzymatic reactions: methylmalonyl-CoA mutase and fumarase. A: methylmalonyl-CoA-mutase catalyzes the isomerization of methylmalonyl-CoA to succinyl-CoA through a radical mechanism. Substitution of a hydrogen in the methyl group (blue) by deuterium leads to a primary kinetic isotope effect, providing evidence for the rate-limiting abstraction of a hydrogen radical from the methyl group at the beginning of the reaction. B: Fumarase catalyzes the hydration of fumarate to malate through a carbenium ion intermediate. A secondary kinetic isotope effect is observed when the hydrogens at C2 or C3 (blue) are substituted with tritium.

QUESTIONS

10.1 What is the approximate rate law for formation of D in the reaction scheme with $k_1 = 100\ s^{-1}$, $k_2 = 1\ s^{-1}$, $k_3 = 10\ s^{-1}$?

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \xrightarrow{k_3} D$$

10.2 Enzyme 1 accelerates the reaction from B to C in the reaction scheme from Question 10.1 by 2-fold, enzyme 2 by 100-fold. What are the approximate integrated rate laws in their presence?

10.3 An inhibitor inhibits the enzyme catalyzing the conversion from B to C in the reaction scheme of Question 10.1 by 2-fold, a different inhibitor achieves a 10-fold inhibition of the first step from A to B. What are the effects on the formation of $[D]$?

10.4 Pyruvate constitutes a metabolic branch point. It can be funneled into the synthesis of alanine with k_1, it can be converted to acetyl-CoA with k_2 and then processed through the citric acid cycle, or it can be used for the multi-step synthesis of valine with k_3, with intermediates for the synthesis of isoleucine and leucine produced along the pathway. What determines the relative flux of pyruvate into the three different pathways? How can energy be made available for the organism during starvation?

10.5 DNA strands (A, B) associate rapidly to mismatched duplexes (I; $k_{A,B\rightarrow I}$) that readily dissociate ($k_{I\rightarrow A,B}$) or slowly rearrange to the base-paired duplex (P) with $k_{I\rightarrow P}$. What is the reaction velocity for duplex formation as a function of the concentration of the DNA strands A and B and what is the apparent rate constant of duplex formation?

REFERENCES

Miller, W. W. and Richards, J. H. (1969) Mechamism of action of coenzyme B12. Hydrogen transfer in the isomerization of methylmalonyl coenzyme A to succinyl coenzyme A. *J Am Chem Soc* **91**(6): 1498–1507.
· kinetic isotope effect to identify the rate-limiting step in the mechanism of methylmalonyl-CoA mutase

Schmidt, D. E. Jr., Nigh, W. G., Tanzer, C. and Richards, J. H. (1969) Secondary isotope effects in the dehydration of malic acid by fumarate hydratase. *J Am Chem Soc* **91**(21): 5849–5854.
· kinetic isotope effect to identify the rate-limiting step in the mechanism of fumarate hydratase

11

Binding reactions

One-step and two-step binding

We will now come back to binding reactions. We have seen one example of a binding event in Scheme 7.3 before, describing the irreversible binding of B to A to form products, such as an AB complex:

$$A + B \xrightarrow{k_1} AB \qquad \text{scheme 11.1}$$

The rate law for this second-order binding reaction is

$$\frac{d[AB]}{dt} = -\frac{d[A]}{dt} = -\frac{d[B]}{dt} = k_1[A][B] \qquad \text{eq. 11.1}$$

We now consider this binding reaction under conditions where the initial concentration of B is much higher than the concentration of A. Under such circumstances, only a small fraction of B will end up forming a complex with A, whereas all A will eventually be bound to B. The concentration of A thus goes from A_0 to zero during the reaction, but the concentration of free B will remain essentially constant. We can therefore approximate the rate law for formation of the AB complex by using the initial concentration of B, B_0, in eq. 11.1:

$$-\frac{d[A]}{dt} \approx k_1[A]B_0 = k_{obs}[A] \qquad \text{eq. 11.2}$$

with

$$k_{obs} = k_1 B_0 \qquad \text{eq. 11.3}$$

The rate constant k_{obs} is the observed rate constant for the exponential law that describes the formation of the AB complex over time. As in previous examples, we now separate the variables for integration.

$$\frac{d[A]}{[A]} = -k_{obs}dt \qquad \text{eq. 11.4}$$

and integrate eq. 11.4 over the concentration range A_0 at $t = 0$ to $[A](t)$ at time t to obtain

$$\ln\left(\frac{[A](t)}{A_0}\right) = -k_{obs}t \qquad \text{eq. 11.5}$$

or

$$[A](t) = A_0\, e^{-k_{obs}t} \qquad \text{eq. 11.6}$$

The rate constant k_{obs} is the observed rate constant for the exponential law that describes the decrease in concentration of A over time. The concentration of the AB complex increases exponentially with k_{obs}:

$$[AB](t) = A_0 - [A](t) = A_0\left(1 - e^{-k_{obs}t}\right) \qquad \text{eq. 11.7}$$

Formation of AB is thus also described by a single exponential function, although complex formation is a second-order reaction. We call this reaction a *pseudo-first-order reaction*, and the conditions under which a reaction behaves as if it were a first-order reaction (here $B_0 \approx [B] \gg A_0$) *pseudo-first-order conditions*. Following the course of a reaction with one partner provided in large excess is a convenient and widely applicable approach to simplify complex reactions and to deduce reaction mechanisms.

Let us now take a look at a binding equilibrium for AB complex formation according to Scheme 11.2.

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AB \qquad \text{scheme 11.2}$$

The rate laws for consumption of A and B are

$$\frac{d[A]}{dt} = -k_1[A][B] + k_{-1}[AB] \qquad \text{eq. 11.8}$$

and

$$\frac{d[B]}{dt} = -k_1[A][B] + k_{-1}[AB] \qquad \text{eq. 11.9}$$

Similarly, we can formulate the rate law in terms of formation of AB as

$$\frac{d[AB]}{dt} = k_1[A][B] - k_{-1}[AB] \qquad \text{eq. 11.10}$$

If we let the reaction proceed under pseudo-first-order conditions, the concentration of B will be constant over the reaction ($[B] \approx B_0$):

$$\frac{d[B]}{dt} \approx 0 \qquad \text{eq. 11.11}$$

and we can simplify the rate laws for A and AB to

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = -k_1[A]B_0 + k_{-1}[AB] \qquad \text{eq. 11.12}$$

and

$$\frac{\mathrm{d}[AB]}{\mathrm{d}t} = k_1[A]B_0 - k_{-1}[AB] \qquad \text{eq. 11.13}$$

According to mass conservation, we can express $[A]$ as A_0–$[AB]$:

$$\frac{\mathrm{d}[AB]}{\mathrm{d}t} = k_1\left(A_0 - [AB]\right)B_0 - k_{-1}[AB] = k_1A_0B_0 + \left(-k_1B_0 - k_{-1}\right)[AB] \qquad \text{eq. 11.14}$$

and can separate the variables

$$\frac{\mathrm{d}[AB]}{k_1A_0B_0 - \left(k_1B_0 + k_{-1}\right)[AB]} = \mathrm{d}t \qquad \text{eq. 11.15}$$

The integral on the left can be evaluated by substitution of u for the denominator, with

$$\frac{\mathrm{d}u}{\mathrm{d}[AB]} = -\left(k_1B_0 + k_{-1}\right)$$

or eq. 11.16

$$\mathrm{d}[AB] = \frac{\mathrm{d}u}{-\left(k_1B_0 + k_{-1}\right)}$$

giving

$$\frac{1}{-\left(k_1B_0 + k_{-1}\right)} \cdot \int \frac{1}{u}\mathrm{d}u = \int_0^t \mathrm{d}t \qquad \text{eq. 11.17}$$

and

$$\frac{1}{-\left(k_1B_0 + k_{-1}\right)} \cdot \ln u = t \qquad \text{eq. 11.18}$$

Substituting back u gives

$$\frac{1}{-\left(k_1B_0 + k_{-1}\right)} \cdot \left[\ln\left(k_1A_0B_0 - \left(k_1B_0 + k_{-1}\right)[AB]\right)\right]_0^{[AB](t)} = t \qquad \text{eq. 11.19}$$

and evaluation of the integral from 0 to $[AB](t)$ then yields

$$\frac{\ln\left(k_1A_0B_0 - \left(k_1B_0 + k_{-1}\right)[AB](t)\right) - \ln\left(k_1A_0B_0\right)}{-\left(k_1B_0 + k_{-1}\right)} = t \qquad \text{eq. 11.20}$$

Eq. 11.20 can be solved for $[AB](t)$ by multiplication with the denominator of the left-hand side and combination of the logarithmic terms remaining, which gives

$$\ln \frac{k_1 A_0 B_0 - (k_1 B_0 + k_{-1})[AB](t)}{k_1 A_0 B_0} = -(k_1 B_0 + k_{-1})t \qquad \text{eq. 11.21}$$

Application of the exponential function then yields

$$\frac{k_1 A_0 B_0 - (k_1 B_0 + k_{-1})[AB](t)}{k_1 A_0 B_0} = e^{-(k_1 B_0 + k_{-1})t} \qquad \text{eq. 11.22}$$

or

$$k_1 A_0 B_0 - (k_1 B_0 + k_{-1})[AB](t) = k_1 A_0 B_0 \cdot e^{-(k_1 B_0 + k_{-1})t} \qquad \text{eq. 11.23}$$

Sign inversion then leads to

$$(k_1 B_0 + k_{-1})[AB](t) = -k_1 A_0 B_0 \cdot e^{-(k_1 B_0 + k_{-1})t} + k_1 A_0 B_0 \qquad \text{eq. 11.24}$$

or

$$(k_1 B_0 + k_{-1})[AB](t) = k_1 A_0 B_0 \left(1 - e^{-(k_1 B_0 + k_{-1})t}\right) \qquad \text{eq. 11.25}$$

which equals

$$[AB](t) = A_0 \frac{k_1 B_0}{k_1 B_0 + k_{-1}} \left(1 - e^{-(k_1 B_0 + k_{-1})t}\right) \qquad \text{eq. 11.26}$$

The exponential term in eq. 11.26 has the form of a simple growth reaction with the observed rate constant

$$k_{\text{obs}} = k_1 B_0 + k_{-1} \qquad \text{eq. 11.27}$$

Thus, the observed rate constant k_{obs} for formation of the AB complex will depend on the concentration of B_0 (in excess), and on the rate constants k_1 and k_{-1} of the elementary reactions. The dependence of the observed rate constant on the concentration B_0 is linear, with a slope of k_1 and y-axis intercept of k_{-1} (Figure 11.1). If we measure the observed rate constant for binding at different concentrations B_0, and then plot k_{obs} as a function of B_0, we can therefore calculate k_1 and k_{-1}, the rate constants for the individual binding and dissociation reactions, by linear regression.

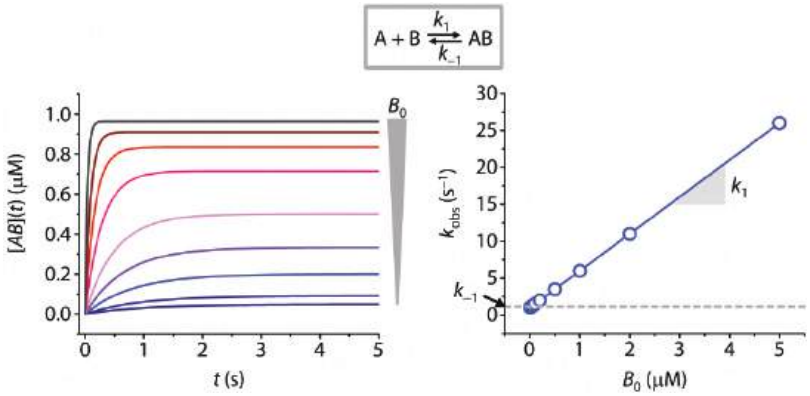


FIGURE 11.1: One-step binding: rate constants. Left: Kinetic transients for complex formation at different concentrations B_0 (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, and 5 μM, from blue to red) under pseudo-first-order conditions ($B_0 \gg A_0$, $[B](t) \approx B_0$). Each transient can be described by a single exponential to extract the observed rate constant k_{obs}. Right: For a binding reaction in a single step, k_{obs} depends linearly on the concentration of B. The rate constants k_1 and k_{-1} can be determined from the slope and the intercept of the line with the y-axis.

The equilibrium constant K_d, the dissociation constant of the AB complex (Section 2.2), can be calculated as k_{-1}/k_1 (eq. 8.21). It is interesting to note that the final concentration of AB in each kinetic transient, which is the concentration of AB in equilibrium at the respective concentrations of B, can also be used to determine the K_d value of the AB complex: a plot of $[AB]_{eq}$ as a function of B_0 represents a binding curve, and the K_d value can be determined using the solution of the quadratic equation that describes the complex concentration in equilibrium as a function of A_0 and B_0 (eq. 11.28, derived in Section 2.2 as eq. 2.25).

$$[AB] = \frac{A_0 + B_0 + K_d}{2} - \sqrt{\left(\frac{A_0 + B_0 + K_d}{2}\right)^2 - A_0 B_0} \qquad \text{eq. 11.28}$$

An agreement of this K_d value with the value calculated from the rate constants k_1 and k_{-1} serves as an internal control for the validity of the applied model (Figure 11.2).

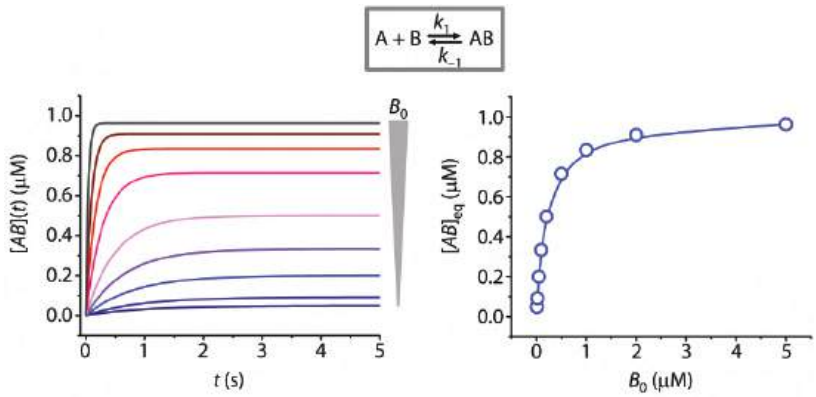


FIGURE 11.2: One-step binding: amplitudes. Left: Kinetic transients for complex formation at different concentrations B_0 (0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, and 5 μM, from blue to red) under pseudo-first-order conditions ($B_0 \gg A_0$, $[B](t) \approx B_0$). Right: The final concentration $[AB]_{eq}$ of each kinetic transient in panel A can be plotted as a function of the concentration B_0 to obtain a binding curve and to determine the K_d value of the AB complex.

Eq. 11.27 only holds if binding is indeed a one-step reaction, and if our approximation ($B_0 \gg A_0$, $[B](t) \approx B_0$) is correct. Should the observed rate constant k_{obs} not show a linear dependence on the concentration B_0, our initial model for binding, *i.e.* binding in one step, will most likely be incorrect, and we have to consider a different model to describe our experimental data.

We can obtain a corresponding expression for the dependence of k_{obs} on the concentration B_0 for a two-step binding reaction, consisting of a rapid equilibrium with the equilibrium dissociation constant K_{d1}, and a second step in which AB is converted to a different complex AB* with the forward and reverse rate constants k_2 and k_{-2} (Scheme 11.3).

$$A + B \underset{K_{d1}}{\rightleftharpoons} AB \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} AB^* \qquad \text{scheme 11.3}$$

K_{d1} for the rapid equilibrium is

$$K_{d1} = \frac{[A][B]}{[AB]} \qquad \text{eq. 11.29}$$

Combined with mass conservation for A and pseudo-first-order conditions ($[B] \approx B_0$), we can write

$$K_{d1} = \frac{\left(A_0 - [AB] - [AB^*]\right)B_0}{[AB]} \qquad \text{eq. 11.30}$$

and solve for $[AB]$

$$[AB] = \frac{A_0B_0 - [AB^*]B_0}{B_0 + K_{d1}} \qquad \text{eq. 11.31}$$

Now we formulate the rate law for product formation as

$$\frac{d[AB^*]}{dt} = k_2[AB] - k_{-2}[AB^*] \qquad \text{eq. 11.32}$$

and substitute $[AB]$ by the expression in eq. 11.31, which gives

$$\frac{d[AB^*]}{dt} = k_2\frac{A_0B_0 - [AB^*]B_0}{B_0 + K_{d1}} - k_{-2}[AB^*] \qquad \text{eq. 11.33}$$

or

$$\frac{d[AB^*]}{dt} = \frac{k_2A_0B_0}{B_0 + K_{d1}} - \frac{k_2B_0}{B_0 + K_{d1}}[AB^*] - k_{-2}[AB^*] \qquad \text{eq. 11.34}$$

which equals

$$\frac{d[AB^*]}{dt} = \frac{k_2A_0B_0}{B_0 + K_{d1}} - \left(\frac{k_2B_0}{B_0 + K_{d1}} + k_{-2}\right)[AB^*] \qquad \text{eq. 11.35}$$

With

$$a = \frac{k_2A_0B_0}{B_0 + K_{d1}} \qquad \text{eq. 11.36}$$

and

$$b = \frac{k_2B_0}{B_0 + K_{d1}} + k_{-2} \qquad \text{eq. 11.37}$$

eq. 11.35 becomes

$$\frac{d[AB^*]}{dt} = a - b[AB^*] \qquad \text{eq. 11.38}$$

or

$$\frac{d[AB^*]}{dt} + b[AB^*] = a \qquad \text{eq. 11.39}$$

By multiplying both sides of eq. 11.39 with e^{bt}, we obtain

$$e^{bt} \cdot \frac{d[AB^*]}{dt} + b \cdot e^{bt}[AB^*] = a \cdot e^{bt} \qquad \text{eq. 11.40}$$

According to the product law (Section 31.6), the left-hand side of eq. 11.40 equals $(uv)'$ with

$$u = [AB^*] \qquad \text{eq. 11.41}$$

and

$$v = e^{bt} \qquad \text{eq. 11.42}$$

We can therefore write eq. 11.40 as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left([AB^*]\cdot e^{bt}\right) = a\cdot e^{bt} \qquad \text{eq. 11.43}$$

Integration then results in

$$\left[e^{bt}\cdot[AB^*]\right]_0^{[AB^*](t)} = a\cdot\left[\frac{1}{b}e^{bt}\right]_0^t \qquad \text{eq. 11.44}$$

which can be evaluated to

$$e^{bt}\cdot[AB^*](t) = a\cdot\left(\frac{1}{b}e^{bt} - \frac{1}{b}\right) = \frac{a}{b}\left(e^{bt}-1\right) \qquad \text{eq. 11.45}$$

and solved for AB*(*t*):

$$[AB^*](t) = \frac{a}{b}\left(1-\frac{1}{e^{bt}}\right) = \frac{a}{b}\left(1-e^{-bt}\right) \qquad \text{eq. 11.46}$$

Substituting *a* and *b* then yields

$$[AB^*](t) = \frac{\dfrac{k_2 A_0 B_0}{B_0 + K_{\mathrm{d1}}}}{\dfrac{k_2 B_0}{B_0 + K_{\mathrm{d1}}} + k_{-2}}\left(1 - e^{-\left(\frac{k_2 B_0}{B_0 + K_{\mathrm{d1}}} + \mathrm{k}_{-2}\right)t}\right) \qquad \text{eq. 11.47}$$

which can be simplified to give the integrated rate law as

$$[AB^*](t) = \frac{k_2 A_0 B_0}{k_2 B_0 + k_{-2}\left(B_0 + K_{\mathrm{d1}}\right)}\cdot\left(1 - e^{-\left(\frac{k_2 B_0}{B_0 + K_{\mathrm{d1}}} + \mathrm{k}_{-2}\right)\mathrm{t}}\right) \qquad \text{eq. 11.48}$$

The observed rate constant k_{obs} for the formation of the AB* complex is then

$$k_{\mathrm{obs}} = k_{-2} + k_2\frac{B_0}{B_0 + K_{\mathrm{d1}}} \qquad \text{eq. 11.49}$$

Thus, in this case the observed rate constant k_{obs} exhibits a hyperbolic dependence on the concentration B_0 (Figure 11.3). We can determine the individual rate constants k_2 and k_{-2}, as well as the equilibrium constant for the first equilibrium, K_{d1}, by measuring the formation of the AB* complex at different concentrations of B. If we keep $B_0 \gg A_0$, then each reaction will be a pseudo-first-order reaction with $[B](t) \approx B_0$, and can be described by a single exponential to calculate k_{obs}. The secondary plot of k_{obs} as a function of B_0 can then be analyzed with eq. 11.49 to obtain K_{d1}, k_2, and k_{-2} (Figure 11.3). We will see in Section 14.3 how we can directly arrive at eq. 11.49 by considering transit times.

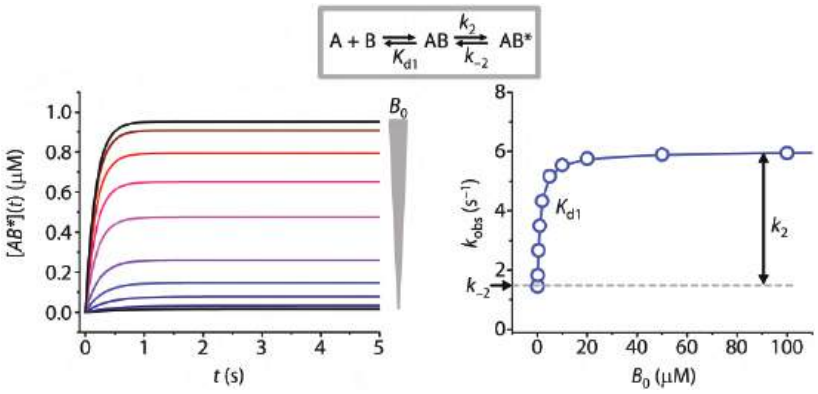


FIGURE 11.3: Two-step binding: rate constants. Left: Kinetic transients for complex formation at different concentrations B_0 (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, and 100 μM from black to blue to red to black) under pseudo-first-order conditions ($B_0 \gg A_0$, $[B](t) \approx B_0$). Each transient can be described by a single exponential to extract k_{obs}. Right: For a two-step binding reaction, k_{obs} depends hyperbolically on the concentration of B. The rate constants k_2 and k_{-2} can be determined from the amplitude of the hyperbola and its intercept with the *y*-axis. The dissociation constant K_{d1} is determined by fitting a hyperbolic curve to the data using eq. 11.49.

As we have seen before for one-step binding, the final concentration of AB* in each kinetic transient corresponds to the equilibrium concentration, and a binding curve is obtained by plotting $[AB^*]_{eq}$ as a function of B_0. From this binding curve, an overall K_d, $K_{d,tot}$, can be determined using the solution of the quadratic equation (eq. 2.25, eq. 11.28). Agreement of this $K_{d,tot}$ with the overall K_d calculated from K_{d1}, k_2, and k_{-2} as

$$K_d = K_{d1}\frac{k_{-2}}{k_2} \qquad \text{eq. 11.50}$$

serves as an internal control for the validity of the applied model (Figure 11.4). Note that measuring $[AB^*]_{eq}$ as a function of B_0 only provides the overall K_d value when the second equilibrium lies on the right-hand side, *i.e.* $K_{d2} < 1$ ($k_2 > k_{-2}$).

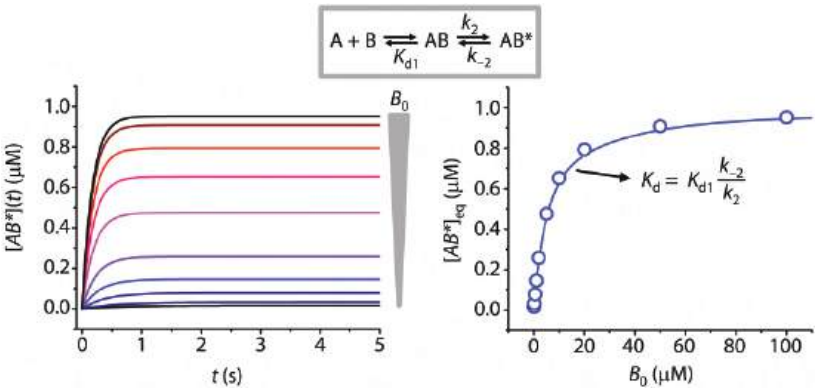


FIGURE 11.4: Two-step binding: amplitudes. Left: Kinetic transients for complex formation at different concentrations B_0 (0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, and 100 μM from black to blue to red to black) under pseudo-first-order conditions ($B_0 \gg A_0$, $[B](t) \approx B_0$). Right: The final value of each kinetic transient in panel A, proportional to $[AB^*]_{eq}$, can be plotted as a function of B_0 to obtain a binding curve. The K_d value of the AB* complex is determined by fitting a curve to the data using eq. 2.25.

We have treated two-step binding under the assumption that we measure the formation of the AB* complex, the end product of the binding reaction. Often, binding reactions are monitored using a spectroscopic probe (see Chapter 28). If formation of the AB complex (Scheme 11.2) does

not lead to a change in signal, *i.e.* AB formation is *spectroscopically silent*, and the change in signal is associated with the conversion of AB to AB*, binding of B under pseudo-first-order conditions follows a single exponential function, and we can describe k_{obs} as a function of B_0 by eq. 11.49. However, the change in signal could instead occur with AB formation, and no further change is associated with the conversion to AB*. In this case, the kinetic transients will follow a double exponential function even under pseudo-first-order conditions. The first, more rapid phase reflects formation of AB from A and B, and the second, slower phase reflects the change in the first equilibrium because of the slow conversion of AB to AB*. If the observed rate constants $k_{obs,1}$ and $k_{obs,2}$ of the two phases are sufficiently different in magnitude, the concentration dependence of $k_{obs,1}$ and $k_{obs,2}$ can be analyzed separately to determine the microscopic rate constants k_1, k_{-1}, k_2, and k_{-2}. If the two phases have similar observed rate constants, their dependence on k_1, k_{-1}, k_2, and k_{-2} is more complex, but it is possible to extract the individual rate constants k_1, k_{-1}, k_2, and k_{-2} from the dependence of the observed rate constants on B_0. For a detailed treatment of these types of reactions and the different cases, the reader is referred to more specialized literature.

QUESTIONS

11.1 The binding of a ligand L to its interaction partner shows a hyperbolic dependence on the ligand concentration under pseudo-first order conditions. k_{obs} as a function of $[L]$ gives K_{d1} = 100 µM and k_2 = 1 s^{-1}. The y-axis intercept is not well defined and k_{-2} cannot be determined accurately from the fit – why? In an equilibrium titration, the overall dissociation constant is determined as $K_{d,overall}$ = 10 nM. Calculate k_{-2}. How can k_{-2} be measured directly and why is the experimental determination of k_{-2} difficult?

11.2 You determine the observed rate constants for binding of an ATPase to a non-hydrolyzable ATP analog under pseudo-first-order conditions. The ATPase concentration is 0.1 µM. k_{obs} shows a linear dependence on the concentration of the ATP analog in a concentration range of 1 – 10 µM. From the slope, you calculate k_+ = 0.1 μM^{-1} s^{-1}, the intercept is k_- = 0.001 s^{-1}. What are the possible meanings of these rate constants?

11.3 Binding of a substrate analog to an enzyme follows a two-step mechanism with $k_{-2} = 1{\cdot}10^{-5}$ s^{-1}. You want to determine the overall K_d value in a fluorescence equilibrium titration. How do you perform the experiment?

11.4 A substance inhibits the activity of an enzyme when added in micromolar concentrations. When you determine the rate constant of binding under pseudo-first-order conditions, you observe a hyperbolic dependence of k_{obs} on the inhibitor concentration with a K_{d1} in the nanomolar range. What can you conclude about k_2 and k_{-2}?

REFERENCES

Bernasconi, C. F. (1976) *Relaxation Kinetics*. New York, Academic Press.
· a detailed book about complex kinetics and the extraction of microscopic rate constants from observed rate constants

12

Single-molecule kinetics

Until now, we have viewed reaction kinetics for synchronized reactions of large ensembles of molecules, meaning reactions where the reactant A is present at a starting concentration of A_0, and we start the conversion of these molecules into product(s) at $t = 0$. From the reaction velocity, *i.e.* the change in concentration of A over time, we have derived integrated rate laws that describe the concentration of A as a function of time. In the following, we will look at reactions and their kinetics from a different point of view: from the perspective of the individual molecules. Single-molecule kinetics become particularly useful for reactions which cannot be triggered. By following the reactions of individual, single molecules, we can derive rate laws and obtain kinetic information for unsynchronized reactions. The methods to follow reactions at the single-molecule level will be treated in Chapters 25 and 27. Here, we will focus on the concepts underlying reactions of single molecules.

12.1 INTEGRATED RATE LAWS, DWELL TIMES, DWELL-TIME AND CUMULATIVE DWELL-TIME DISTRIBUTIONS

Before we dive into single-molecule kinetics of more complex reactions, let us first recapitulate what happens when we analyze a synchronized, unimolecular reaction, in which an ensemble of reactant molecules is converted into products (Figure 12.1). We start this reaction by providing the reactant A, and follow its conversion to the product B using a suitable probe. The integrated rate law for such a reaction, derived in Section 7, is a single-exponential function of the form

$$[A](t) = A_0 e^{-kt} \qquad \text{eq. 7.7}$$

where A_0 is the starting concentration, and k is the rate constant for the reaction of A to product. How does the exponential decrease in concentration that we observe relate to the behavior of the individual molecules of A? On a molecular level, the decreasing concentration is a measure for the remaining molecules of A, those that have not yet reacted at time t. The (negative) slope of the exponential function, the (negative) first derivative (eq. 12.1), then is a measure for the molecules that are reacting as a function of time (Figure 12.1). The negative sign in eq. 12.1 stems from the fact that we

now count the numbers of molecules that have reacted, whereas the integrated rate law counts the molecules that are still present.

$$-\frac{\mathrm{d}[A]}{\mathrm{d}t} = kA_0e^{-kt} \qquad \text{eq. 12.1}$$

The number of molecules reacting at time t in turn is connected to the reaction velocity at time t (Chapter 6). The reaction velocity (see Figure 6.1) and the number of molecules reacting (Figure 12.1) are not constant, but decrease with time. From this behavior, we can imply that not all molecules react at the same time: many react at the beginning, fewer react at later time points. Hence, if we watched every single molecule A in the reaction mixture as a function of time, we would observe that some convert to B right away, while others take considerably longer (Figure 12.1). The time until a molecule reacts is termed its *waiting time* or *dwell time*. The negative first derivative of the integrated rate law (eq. 12.1) is the *dwell-time distribution*. If we want to count all molecules that have reacted at time t, we have to calculate the difference between the current concentration of A and the starting concentration A_0 (Figure 12.1). The resulting function is the integrated form of the dwell-time distribution, termed the *cumulative dwell-time distribution*.

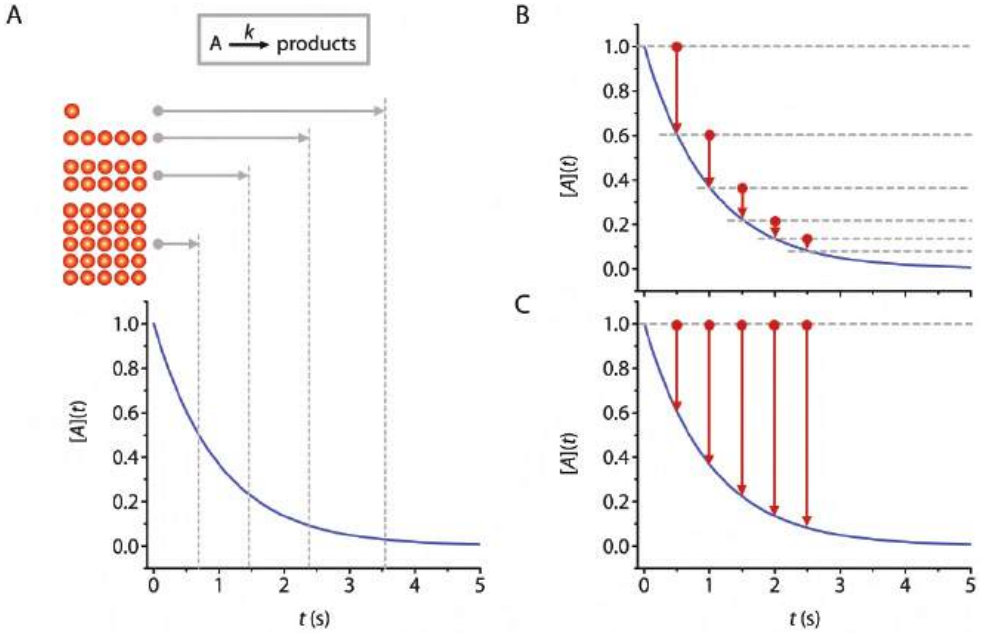


FIGURE 12.1: Relationship between integrated rate law, dwell-time, and cumulative-dwell time distributions. A: Integrated rate law (blue) for a unimolecular reaction A→B, and the times individual molecules (orange) spend before converting into product B. Many molecules react with small dwell times, few molecules have large dwell times. B: Relation of the integrated rate law (blue) and its (negative) first derivative, the dwell-time distribution. The dwell-time distribution reflects the number of molecules that are reacting at time t (red arrows). C: Relation of the integrated rate law (blue) and the cumulative dwell-time distribution. The cumulative dwell-time distribution corresponds to the number of all molecules that have reacted until time point t (red arrows).

12.2 SINGLE-MOLECULE KINETICS AND THE PRINCIPLE OF ERGODICITY

We have introduced the concept of ergodicity briefly in Chapter 3 in the context of statistical thermodynamics (Section 3.1). There, we stated that the time an individual molecule of a large ensemble of molecules spends in each possible state over time recapitulates the distribution of the large ensemble of molecules over these states at any given point in time. This principle rationalizes why we can obtain information on rate constants of reactions not only from analyzing synchronized reactions by observing the collective behavior of ensembles of molecules over time, but also from

unsynchronized reactions, by observing the temporal behavior of individual molecules: if we watch a single molecule for a sufficiently long time, it will sample all possible states with the same probability as the numerous molecules of the ensemble at a particular time point (Figure 12.2). In other words, the relative time a single molecule spends in each state is equivalent to the relative number of molecules that adopt this particular state at a defined time point. Again, the temporal average for the behavior of the single molecule contains the equivalent information as the average over the population of molecules in the ensemble. Note that the principle of ergodicity only holds for averaging over large numbers, either over large numbers of molecules in the ensemble or over a large number of time points for a single molecule.

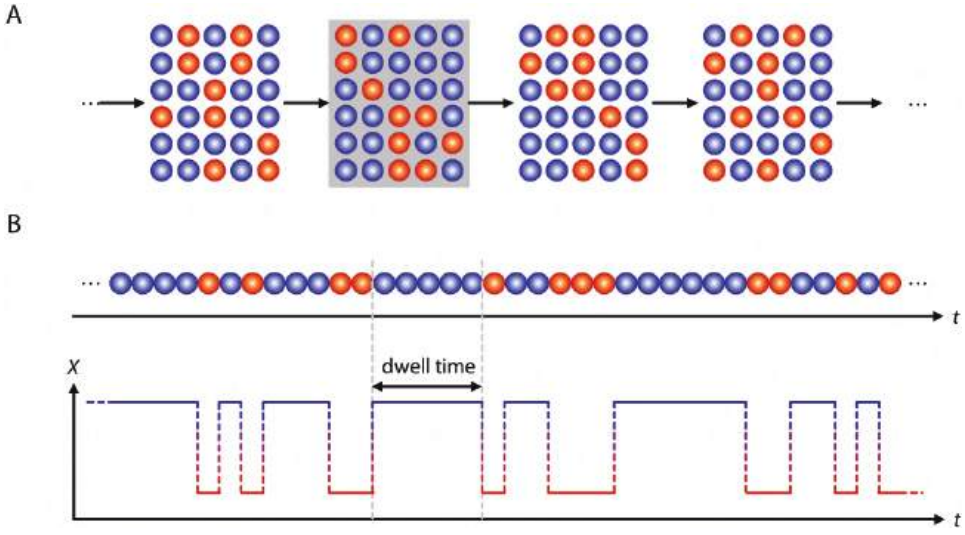


FIGURE 12.2: The principle of ergodicity: ensemble and single-molecule behavior. A: An ensemble of molecules in equilibrium that interconvert between states A and B (blue and red). B: Behavior of a single molecule of the ensemble in A observed as a function of time (top). The time the molecule spends in each state reflects the distribution over both states in the ensemble in A at a certain time point (gray rectangle). Switching of this molecule between states A and B can be followed as a function of time using a suitable probe (*X*; bottom) that provides different signals for A and B. The dwell time is the time between two transitions, *i.e.* the time the molecule spends in a particular state between transitions.

We now have different ways of deriving rate constants for reactions from observing single molecules. For irreversible reactions, we can observe individual molecules of A separately, and can measure the time when each molecule converts from A to B (Figure 12.1). This is the only way to derive rate constants for irreversible reactions on a single-molecule level; in this case, we are still limited to synchronized reactions started by a triggering event at $t = 0$. For reversible reactions, where molecules convert from A to B and back to A with specific rate constants, we can determine these rate constants on a single-molecule level by observing a single molecule over (ideally infinite) time, alternating between states A and B (Figure 12.2). From experimental data, we can collect all dwell times in states A and B in discrete time windows, meaning we count how many molecules have reacted in a first time window of a certain width Δt, how many molecules have reacted with a dwell time in the subsequent time window of equal duration, and so on (Figure 12.1, see Box 25.9). The time window is also termed *bin*, the width Δt of the time window is the *bin size*, also denoted as t_{bin}. The resulting histogram of dwell times is the *dwell-time histogram*, a discrete representation of the continuous dwell-time distribution (eq. 12.1). The value in each bin of the dwell-time histogram is the difference Δn_{A} between the number of molecules n_{A} still present at the beginning of the bin, and the number of molecules still present at the end of the bin, plotted at time t_k, the center of the bin:

$$\Delta n_{\text{A}}(t_{\text{k}}) = n_{\text{A}}\left(t_{\text{k}} - \frac{t_{\text{bin}}}{2}\right) - n_{\text{A}}\left(t_{\text{k}} + \frac{t_{\text{bin}}}{2}\right) \qquad \text{eq. 12.2}$$

The values for t_k are $kt_{bin}+t_{bin}/2$, where k is an integer from 0 to the total number of bins. For $t_{bin}\rightarrow 0$, eq. 12.2 becomes

$$\lim_{t_{bin}\rightarrow 0} \Delta n_A(t) = kA_0 e^{-kt} = -\frac{d[A]}{dt} \qquad \text{eq. 12.3}$$

Thus, the dwell-time histogram is a binned version of the negative first derivative of the integrated rate law (Figure 12.1), and can be described by a single-exponential function with the same rate constant as the integrated rate law. Note that, due to the calculation of the first derivative, the amplitude of a dwell-time histogram is determined not only by the starting concentration of the reactant, A_0, but also by the rate constant k of the reaction!

As an alternative to the dwell-time histogram, we can also look at the reaction by counting all molecules that have reacted when a certain time t has passed, and generate a *cumulative dwell-time histogram*, the binned form of the cumulative dwell-time distribution (Figure 12.1):

$$n_A(t_k) = \sum_{k=0}^{n} n_A\left(t_k - \frac{t_{bin}}{2}\right) - n_A\left(t_k + \frac{t_{bin}}{2}\right) \qquad \text{eq. 12.4}$$

where n denotes the total number of bins. For $t_{bin}\rightarrow 0$, the cumulative dwell-time histogram becomes the integral of the dwell-time histogram:

$$n_A(t) = \int_0^t kA_0 e^{-kt} dt = A_0\left(1 - e^{-kt}\right) \qquad \text{eq. 12.5}$$

Here, $n_A(t)$ denotes all molecules that have reacted between the time t = 0 to the actual time t. According to eq. 12.5, the amplitude of the cumulative dwell-time histogram depends on the starting concentration A_0, but not on the rate constant of the reaction. This is an important difference to the dwell-time histogram, and makes the cumulative dwell-time histogram comparable to the integrated rate law. However, eq. 12.5 is not identical to the exponential decrease in the integrated rate law, but equals $A_0 - [A](t)$, as it counts the molecules that have reacted, not those that are still present. Understanding the relationship between dwell-time and cumulative dwell-time histograms, and their relation to the integrated rate law describing the corresponding ensemble reaction, is key for the interpretation of these histograms.

12.3 SINGLE-MOLECULE KINETICS OF REVERSIBLE REACTIONS

Now that we understand the concept of dwell times and the principles of reactions at the molecular level, we can apply our knowledge to reversible reactions, such as an interconversion of A and B (Figure 12.2; see Sections 8.1 and 9.1):

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B \qquad \text{scheme 8.1}$$

We have seen in Section 8.1 that this reaction is described by the integrated rate laws

$$[A](t) = \frac{A_0}{k_1 + k_{-1}}\left(k_{-1} + k_1 e^{-(k_1 + k_{-1})t}\right) \qquad \text{eq. 8.17}$$

and

$$[B](t) = \frac{A_0}{k_1 + k_{-1}} k_1\left(1 - e^{-(k_1 + k_{-1})t}\right) \qquad \text{eq. 8.18}$$

For a synchronized reaction, we start from a concentration A_0, with $B_0 = 0$, and observe how the reaction approaches the equilibrium concentrations $[A]_{eq}$ and $[B]_{eq}$, with an apparent rate constant that is the sum of the individual rate constants k_1 and k_{-1} (Figure 12.3). If we view this reaction on a molecular level, we can again watch all individual molecules of A and stop the time until they have reacted to B. However, in this case, any molecule that has converted to B can again return to A, complicating the analysis. Instead of watching every single molecule of A from the beginning and stopping the time until it converts to B, we now follow the spontaneous interconversion of individual molecules from A to B and back to A in the dynamic equilibrium, for as long as possible with our experimental approach. Thereby, we register all transitions from A to B and from B to A over time, and can histogram the dwell times of the molecules in states A and in B. In this way, we can analyze the reversible reactions as two (irreversible) half-reactions, in which a state A converts into state B with a rate constant k_1 (forward reaction), and B converts into A with a rate constant k_{-1} (reverse reaction; Figure 12.3). Note that the dwell times of the molecules in states A and B characterize the decay of the respective species, and thus the reaction away from this state. The kinetic information, *i.e.* the rate constant for the underlying process, is again encoded in the distribution of these dwell times. The irreversible half-reactions for this system are described by very simple integrated rate laws in the form of a single-exponential decay:

$$[A](t) = A_0 e^{-k_1 t} \qquad \text{eq. 12.6}$$

and

$$[B](t) = B_0 e^{-k_{-1} t} \qquad \text{eq. 12.7}$$

The dwell-time distributions for A and B, analyzed separately for each of the two states by collecting all dwell times measured for the transitions from A to B and from B to A, characterize only the forward or only the reverse reaction. These dwell-time histograms correspond to the negative first derivative of the integrated rate laws in eq. 12.6 and eq. 12.7, and follow single exponentials of the form

$$\Delta n_A(t_k) = n_A\left(t_k - \frac{t_{bin}}{2}\right) - n_A\left(t_k + \frac{t_{bin}}{2}\right) \qquad \text{eq. 12.2}$$

which, for $t_{bin} \to 0$ becomes

$$\lim_{t_{bin}\to 0} \Delta n_A(t) = \frac{d[A]}{dt} = k_1 A_0 e^{-k_1 t} \qquad \text{eq. 12.8}$$

and

$$\Delta n_B(t_k) = n_B\left(t_k - \frac{t_{bin}}{2}\right) - n_B\left(t_k + \frac{t_{bin}}{2}\right) \qquad \text{eq. 12.9}$$

or

$$\lim_{t_{bin}\to 0} \Delta n_B(t) = \frac{d[B]}{dt} = k_{-1} A_B e^{-k_{-1} t} \qquad \text{eq. 12.10}$$

Correspondingly, the cumulative dwell-time histograms are described by

$$n_A(t) = \int_0^t k_1 A_0 e^{-k_1 t} dt = A_0\left(1 - e^{-k_1 t}\right) = A_0 - [A](t) \qquad \text{eq. 12.11}$$

and

$$n_B(t) = \int_0^t k_{-1} B_0 e^{-k_{-1}t} dt = B_0\left(1 - e^{-k_{-1}t}\right) = B_0 - [B](t) \qquad \text{eq. 12.12}$$

By analysis of the (cumulative) dwell-time histograms, we can thus directly determine the individual rate constants k_1 and k_{-1}. This is different from the ensemble approach, where the integrated rate law only provides an apparent rate constant that is the sum of k_1 and k_{-1}, and where we have to calculate the values of k_1 and k_{-1} from the amplitude information (Section 8.1).

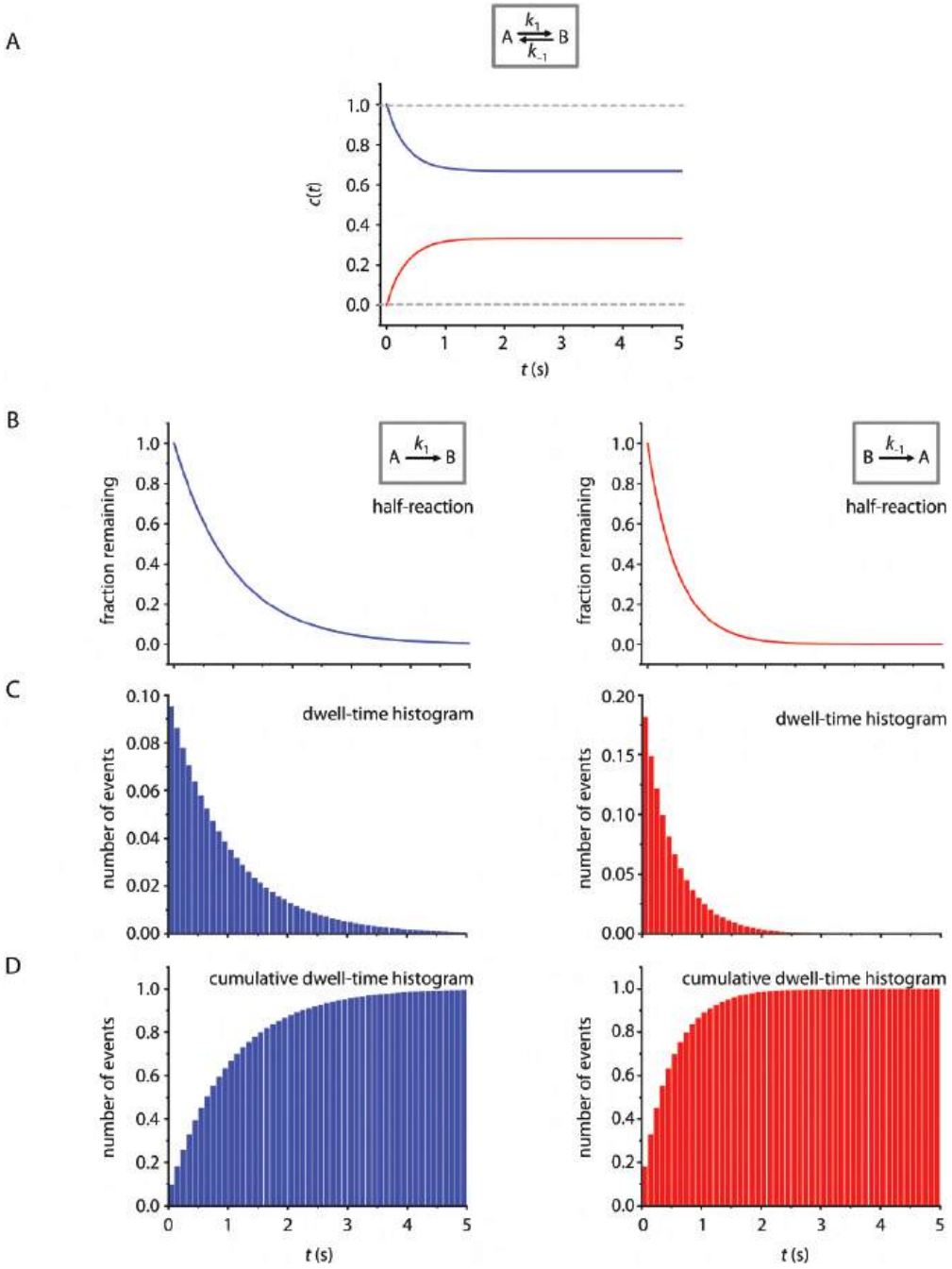


FIGURE 12.3: Integrated rate law, dwell-time, and cumulative-dwell time histograms for reversible reactions. A: Integrated rate laws [*A*](*t*), [*B*](*t*) (red, blue) for a reversible reaction. B: Half-reactions studied in single-molecule kinetics. C: Dwell-time histograms: D: Cumulative dwell-time histograms.

12.4 THE LIMITS OF DWELL-TIME HISTOGRAMS: POTENTIAL PITFALLS

In single-molecule kinetics, large numbers of observations are not only required to ensure the validity of the ergodic principle, but also to obtain well-resolved, smooth dwell-time histograms that unambiguously define the underlying exponential. When the dwell times determined from the temporal behavior of single molecules are collected into histograms, they need to be binned into suitable intervals. The number of dwell times determined in an experiment is always limited, and, as a consequence, the optimal bin size is a compromise (Figure 12.4). It should be small enough to reveal the underlying exponential function but large enough to ensure a reasonable number of events per bin. The smallest bin size possible is determined by the time resolution of the experiment: if we measure the signal serving as a probe for the different states in intervals t_{int}, then this is the smallest bin size possible; we cannot use bin sizes below that. However, using the smallest bin size possible to construct dwell-time histograms often is not the best choice. Depending on the number of dwell times collected, dwell-time histograms with such small bin sizes will contain very few events in each bin. As a consequence, the dwell-time distribution will appear rather rugged, making it difficult to describe it with an exponential function to extract the rate constant. There are two ways to amend this problem: collect more dwell times or make bins larger to collect more events per bin. However, the bin size also limits the rate constants we can resolve. Therefore, the maximum bin size possible depends on the rate constant of the reaction observed. To determine a rate constant k, the bin size of the dwell-time histogram should be divided into at least three bins per lifetime τ (*i.e.* the inverse of the rate constant). For example, the dwell times for a reaction with $k = 1\ s^{-1}$ ($\tau = 1$ s) should be collected in histograms with a bin size t_{bin} of not more than 0.33 s. The other way round, measurements that require bin sizes of 0.5 s to generate smooth histograms can only resolve reactions with lifetimes of $\tau > 1.5$ s, or rate constants of $k < 0.67\ s^{-1}$. To sample the entire exponential distribution of dwell times, it is also important to collect dwell times that are sufficiently large to define the trailing end of the exponential function. This is achieved when dwell times up to at least three times the lifetime are included.

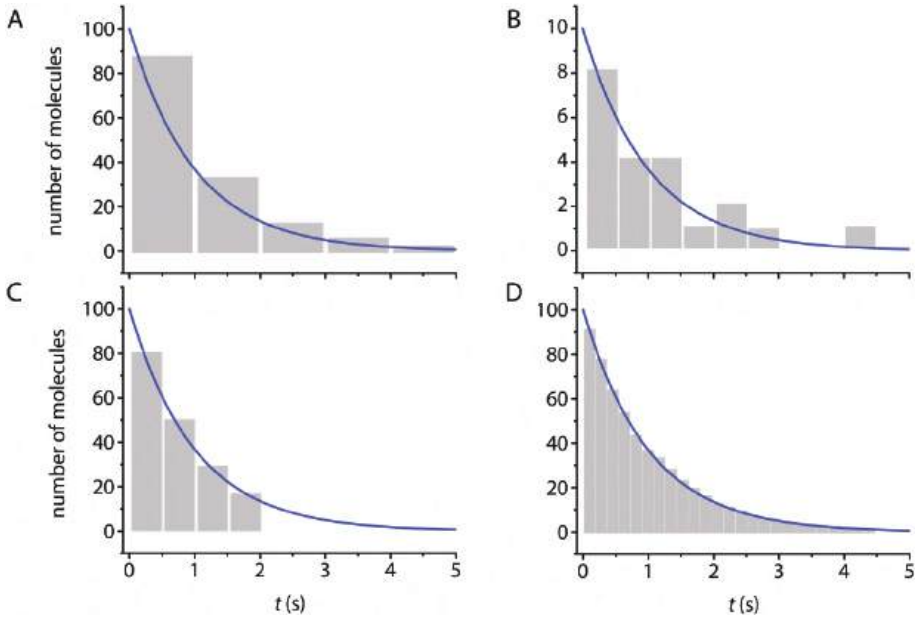


FIGURE 12.4: Effect of binning on dwell-time histograms. A: Too few bins to sample the exponential function and resolve the rate constant. B: Too few events per bin. C: Longer dwell times not collected. D: Well-defined dwell-time histogram. The blue curve is a single-exponential function.

12.5 DOUBLE-EXPONENTIAL DWELL-TIME HISTOGRAMS

In certain cases, dwell-time histograms are not single exponentials, but have two or more phases. For example, this can be the case for parallel or sequential reactions (see Chapter 8). Double-exponential dwell-time histograms come with specific challenges in data analysis.

The first complication is associated with the binning process (Figure 12.5): to resolve two significantly different rate constants, the minimum number of bins required is substantially different. As an illustrative example, we consider a biphasic reaction, with a fast phase governed by a rate constant of $k_1 = 1\ s^{-1}$ ($\tau_1 = 1$ s), and a slow phase governed by $k_2 = 0.1\ s^{-1}$ ($\tau_2 = 10$ s). To resolve the fast phase, we need a dwell-time histogram with a bin size of approx. 0.3 s, collecting dwell times up to 3 s. For the second phase, a bin size of 3 s would be sufficient, but we need dwell times up to 30 s to define the tail of the exponential. What happens if we use a bin size of 3 s to generate the dwell-time histogram? In this case, the slow phase is well resolved by three bins per lifetime. However, the events belonging to the fast phase are not resolved: most of the short dwell times representing this phase are collected in the first bin (Figure 12.5). If, on the other hand, we construct the dwell-time histogram with the smaller bin size of 0.3 s, we would nicely resolve the fast phase, but would distribute the events from the slow phase over too many bins to obtain a smooth histogram and determine the underlying rate constant (Figure 12.5). The binning problem can be overcome by analyzing dwell-time histograms with different bin sizes. This is particularly important whenever a suspiciously high number of dwell times ends up in an over-proportionately large first bin. A second possibility is the construction of cumulative dwell-time histograms instead of dwell-time histograms. As cumulative dwell-time histograms accumulate all events with dwell times up to a certain value in each bin, they naturally contain a larger number of events per bin than dwell-time histograms with the same bin size.

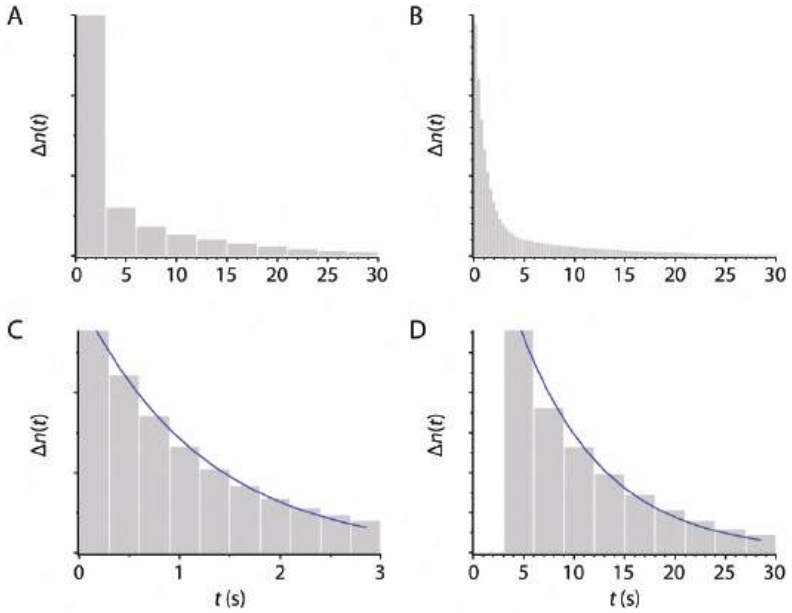


FIGURE 12.5: Binning dwell times of double-exponential reactions. Dwell-time histograms for a biphasic reaction with $A_0 = 1$ μM, $k_1 = 1\ s^{-1}$, $k_2 = 0.1\ s^{-1}$, and $x = 0.5$. A: Bin size 3 s. Most events of the fast phase are summarized in the first bin. B: Bin size 0.3 s. The fast phase is resolved. C: First 3 s of the dwell-time histogram with $t_{bin} = 0.3$ s. The blue line is a single-exponential with a rate constant of $k_1 = 1\ s^{-1}$. D: Dwell-time histogram with a bin size of 3 s without first bin. The blue line is a single-exponential with a rate constant of $k_2 = 0.1\ s^{-1}$.

The second complication is related to the amplitude information of dwell-time histograms. We have seen that dwell-time histograms constitute the negative first derivative of integrated rate laws (Section 12.1). The integrated rate law for a double-exponential reaction is

$$[A](t) = xA_0e^{-k_1t} + (1-x)A_0e^{-k_2t} \qquad \text{eq. 12.13}$$

where x describes the fraction of the molecules of A reacting with k_1, and $(1-x)$ the fraction reacting with k_2. The negative first derivative describing the dwell-time histogram is then

$$\Delta n_A(t) = -\frac{dA}{dt} = xk_1A_0e^{-k_1t} + (1-x)k_2A_0e^{-k_2t} \qquad \text{eq. 12.14}$$

The same way as we have seen for single-exponential dwell-time histograms, the rate constants k_1 and k_2 of a biphasic reaction can be determined directly from dwell-time histograms by describing them with a double-exponential function. When analyzing the reaction amplitudes of the two phases, however, we need to take into account that the amplitudes of the dwell-time histogram become weighted by the rate constants for the two processes (Figure 12.6). When k_1 and k_2 are similar, the resulting effect is small, but if the two processes occur on different time scales, then the effect of this weighting needs to be considered in an analysis of the amplitudes. Again, this problem is circumvented by the analysis of cumulative dwell-time histograms, where the amplitudes are not weighted by the rate constants (Figure 12.6). For the biphasic reaction described by eq. 12.13, the cumulative dwell-time histogram is obtained by integration of eq. 12.14:

$$n_A(t) = A_0\left(1 - xe^{-k_1t} + (1-x)e^{-k_2t}\right) \qquad \text{eq. 12.15}$$

Analysis of the cumulative well-time histogram with a double-exponential function thus retrieves both the rate constants and the amplitudes of the two phases directly.

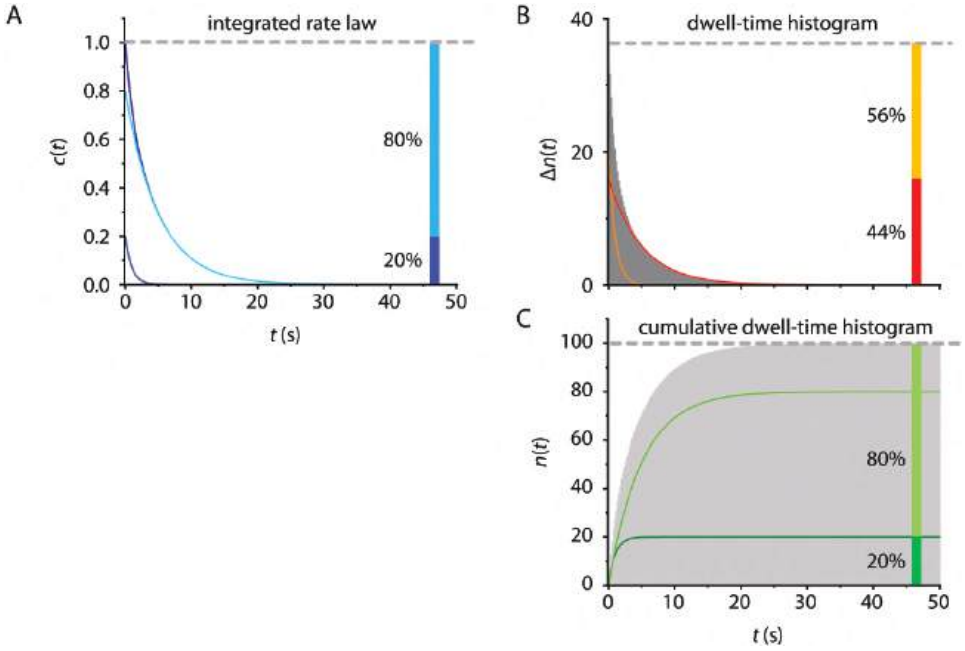


FIGURE 12.6: Binning dwell times of double-exponential reactions. (A) Integrated rate laws for a biphasic reaction with $k_1 = 1\ s^{-1}$, $k_2 = 0.2\ s^{-1}$, $A_1 = 0.2$ and $A_2 = 0.8$. B: Dwell-time histogram of the same biphasic reaction. C: Cumulative dwell-time histogram. While the amplitudes are weighted with the rate constants in the dwell-time histogram, the cumulative dwell-time histogram reflects the true amplitudes of the two phases.

Single-molecule approaches are particularly informative for systems where reactions cannot be triggered and synchronized. This is often the case for enzymes undergoing conformational cycles. In this case, the spontaneous fluctuations of individual, unsynchronized molecules can simply be followed as a function of time. The time traces directly reveal the sequence in which different states are populated, *i.e.* the connectivity of these states. In addition, the times molecules spend in each state provide information on the rate constants for these transitions. We will revisit single-molecule approaches in Chapter 25, especially in Section 25.7 on single-molecule fluorescence microscopy, and in Chapter 27, particularly in Sections 27.2.2–27.2.4 on single-molecule force measurements.

QUESTIONS

12.1 What is a suitable bin size to determine rate constants $k = 0.3\ s^{-1}$ or $k = 30\ s^{-1}$ from a dwell-time histogram? Which range of dwell times should be collected?

12.2 A biphasic reaction has a fast phase with $k_1 = 10\ s^{-1}$, and a slower phase with $k_2 = 0.2\ s^{-1}$. A dwell-time histogram is generated with a bin size of $t_{bin} = 0.1$ s. Calculate the fraction of events from the slow and fast phase in the first bin. How does this change if the bin size is increased to 1 s?

12.3 A double-exponential reaction occurs with $k_1 = 1\ s^{-1}$ and $k_2 = 0.2\ s^{-1}$. The amplitudes of the two phases are $A_1 = 0.2$ and $A_2 = 0.8$ ($A_1 = 0.5$ and $A_2 = 0.5$). What are the amplitudes of the two phases in a dwell-time histogram?

12.4 Show for a unimolecular reaction that the mean dwell time equals the lifetime of the reaction.

12.5 Show that the dwell-time histogram approaches the continuous dwell-time distribution for small bin sizes. Hint: Use the l'Hôpital's rule.

REFERENCES

Andreou, A. Z. and Klostermeier, D. (2012) Conformational changes of DEAD-Box helicases monitored by single molecule fluorescence resonance energy transfer. *Methods Enzymol* **511**: 75–109.

Bagshaw, C. R. (2017) *Biomolecular Kinetics – A Step-by-Step Guide*. New York, CRC Press.
· a kinetics textbook with a dedicated chapter on single-molecule kinetics

Blanco, M. and Walter, N. G. (2010) Analysis of complex single-molecule FRET time trajectories. *Methods Enzymol* **472**: 153–178.
· review discussing the analysis of dwell times from single-molecule data

Chakraborty, A., Krause, L. and Klostermeier, D. (2022) Determination of rate constants for conformational changes of RNA helicases by single-molecule FRET TIRF microscopy. *Methods* **204**: 428–441.
· review describing the determination of rate constants for conformational changes from single-molecule FRET data

Klostermeier, D. (2022) Analysis of the conformational space and dynamics of RNA helicases by single-molecule FRET in solution and on surfaces. *Methods Enzymol* **673**: 251–310.
· review describing the extermination of rate constants for conformational changes from single-molecule FRET data

Lee, T. H. (2009) Extracting kinetics information from single-molecule fluorescence resonance energy transfer data using hidden Markov models. *J Phys Chem B* **113**(33): 11535–11542.
· review explaining hidden Markov modeling and its application to rate constant determination from single-molecule FRET data

Li, J., Zhang, L., Johnson-Buck, A. and Walter, N. G. (2020) Automatic classification and segmentation of single-molecule fluorescence time traces with deep learning. *Nat Commun* **11**(1): 5833–5843.
· a neural-network approach to select single-molecule time traces for kinetic analyses

McKinney, S. A., Joo, C. and Ha, T. (2006) Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys J* **91**(5): 1941–1951.
· review explaining hidden Markov modeling and its application to rate constant determination from single-molecule FRET data

Schmid, S., Gotz, M. and Hugel, T. (2016) Single-molecule analysis beyond Dwell times: Demonstration and assessment in and out of equilibrium. *Biophys J* **111**(7): 1375–1384.
· a maximum-likelihood approach to derive kinetic models from single-molecule data

Thomsen, J., Sletfjerding, M. B., Jensen, S. B., Stella, S., Paul, B., …, Hatzakis, N. S. (2020) DeepFRET, a software for rapid and automated single-molecule FRET data classification using deep learning. *Elife* **9**: e60404.
· a deep-learning approach to select single-molecule time traces for kinetic analyses

13

Steady-state (enzyme) kinetics

In the previous Chapters we have treated reaction velocities where we explicitly took into account changes in all reactants over time. We have seen that we can create pseudo-first-order conditions by providing one reactant in large excess, such that its concentration remains constant during the reaction (Chapter 11). We will now derive the quantitative concepts underlying *steady-state reactions*, *i.e.* reactions where one component is in large excess, such that its concentration does not change with time during the reaction, and the reaction continues without reaching an endpoint. This treatment is particularly useful for enzyme-catalyzed reactions, where the substrate is usually in large excess.

An enzymatic reaction starts with binding of the substrate S to the enzyme E, leading to the formation of an enzyme-substrate complex ES. The equilibrium dissociation constant of the ES complex is K_S. The ES complex is converted into an enzyme-product complex EP in a first-order reaction with the rate constant k_p. Finally, the EP complex dissociates into enzyme and products, characterized by the equilibrium dissociation constant K_P (Scheme 13.1). The enzyme is regenerated at the end of the reaction, in agreement with its action as a catalyst of the reaction from S to P (see Chapter 16).

$$E + S \underset{K_S}{\rightleftharpoons} ES \xrightarrow{k_p} EP \overset{K_P}{\rightleftharpoons} E + P \qquad \text{scheme 13.1}$$

We now simplify this scheme by assuming that there is only one central complex, namely ES, that directly decomposes into enzyme and product(s) in an irreversible reaction with the rate constant k_p. Scheme 13.1 then becomes

$$E + S \underset{K_S}{\rightleftharpoons} ES \xrightarrow{k_p} E + P \qquad \text{scheme 13.2}$$

We can also specify the rate constants for the forward and reverse reaction of the first equilibrium:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_p} E + P \qquad \text{scheme 13.3}$$

Now we can derive rate laws for product formation under two different conditions. We will first treat the case where formation of the ES complex is rapid, and the rate of product formation is rate-limiting, *i.e.* k_1, $k_{-1} > k_p$ (the *rapid equilibrium approach*). After that, we derive the rate law for the case where product formation is not rate-limiting, but occurs on a comparable time scale to the equilibration of E and ES.

13.1 RAPID EQUILIBRIUM (MICHAELIS-MENTEN FORMALISM)

If the enzyme-substrate complex ES is formed in a rapid equilibrium, we can define the dissociation constant of the ES complex, K_S, as

$$K_S = \frac{[E][S]}{[ES]} \qquad \text{eq. 13.1}$$

and obtain

$$[ES] = \frac{[E][S]}{K_S} \qquad \text{eq. 13.2}$$

for the concentration of ES in equilibrium.
The velocity of product formation v is determined by the rate-limiting step, and is

$$v = k_p[ES] \qquad \text{eq. 13.3}$$

The *maximum velocity* v_{max} is reached if all enzyme molecules are bound to a substrate, or, put in other words, under saturation of the enzyme with substrate. In this case, we can substitute $[ES]$ by the total enzyme concentration $[E]_t$.

$$v_{max} = k_p[E]_t \qquad \text{eq. 13.4}$$

We can now divide eq. 13.3 by $[E]_t$. On the right-hand side, we write $[E]_t$ as the sum of the concentrations of the two possible enzyme forms, E (free form) and ES (substrate-bound form).

$$\frac{v}{[E]_t} = \frac{k_p[ES]}{[E]+[ES]} \qquad \text{eq. 13.5}$$

Note that eq. 13.5 relates the kinetic behavior of the entire system to the fraction of active state $[ES]$ in relation to all accessible different states, in this case $[E]$ and $[ES]$ of the enzyme (see Sections 3.2, 3.4). We can now substitute $[ES]$ by the expression in eq. 13.2,

$$\frac{v}{[E]_t} = \frac{k_p \frac{[E][S]}{K_S}}{[E]+\frac{[E][S]}{K_S}} \qquad \text{eq. 13.6}$$

Division by k_p and canceling $[E]$ yields

$$\frac{v}{k_p[E]_t} = \frac{\frac{[S]}{K_S}}{1+\frac{[S]}{K_S}} \qquad \text{eq. 13.7}$$

The denominator on the left-hand side of the equation above denotes the maximum reaction velocity v_{max} (eq. 13.4). The fraction on the right-hand side defines the fraction of enzyme with their active

site occupied by substrate, or the probability to find the system in the bound state (see Section 3.1). Note that the expression is equivalent to the expressions derived from statistical thermodynamics, with the free enzyme E as the reference state (see Section 3.2). By expanding eq. 13.7 with K_S on the right, we obtain the *rate* or *velocity equation*

$$\frac{v}{v_{max}} = \frac{[S]}{K_S + [S]} \qquad \text{eq. 13.8}$$

In this form, the velocity equation relates the actual velocity v at a defined concentration of substrate, where only some of the active sites available are occupied with substrate, to the maximum velocity v_{max}, achieved at a substrate concentration where all active sites are occupied. The ratio v/v_{max} thus corresponds to the fractional saturation of the substrate binding sites (see Section 3.4, 16.1).

Often, velocity equations for enzyme-catalyzed reactions are written in the form

$$v = v_{max} \cdot \frac{[S]}{K_S + [S]} \qquad \text{eq. 13.9}$$

Both eq. 13.8 and eq. 13.9 describe a hyperbola (Figure 13.1). For $[S] = K_S$, the half-maximum velocity is reached, and v/v_{max} is 0.5. For $[S] \rightarrow \infty$, *i.e.* saturation of the enzyme with substrate, the velocity reaches v_{max}; the ratio v/v_{max} reaches a value of unity. The constant K_S is often referred to as the *Michaelis-Menten constant* K_M. We will use K_S to refer to the true equilibrium constant of ES dissociation, and K_M for apparent equilibrium dissociation constants.

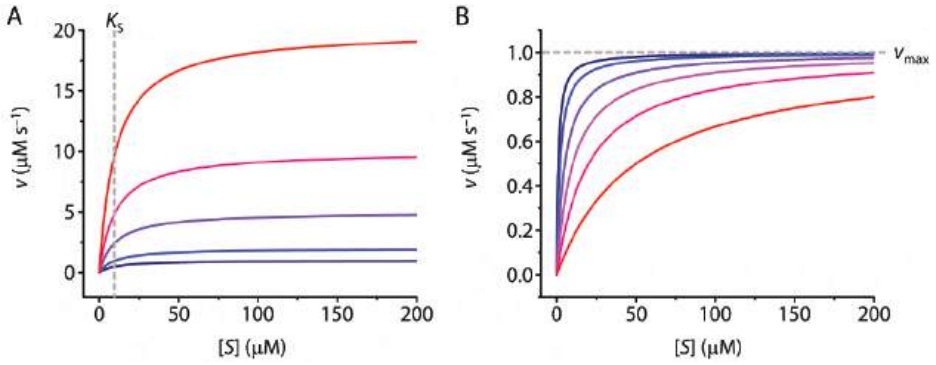


FIGURE 13.1: Reaction velocity as a function of substrate concentration. A: Hyperbolic dependence of the reaction velocity v on the free substrate concentration [S] for v_{max} = 1 μM s^{-1}, 2 μM s^{-1}, 5 μM s^{-1}, 10 μM s^{-1}, and 20 μM s^{-1} (blue to red) and K_S = 10 μM. B: Hyperbolic dependence of the reaction velocity v on the free substrate concentration [S] for v_{max} = 1 μM s^{-1} and K_S = 1 μM, 2 μM, 5 μM, 10 μM, and 20 μM (blue to red). The Michaelis-Menten constant K_S is the substrate concentration for which v is 0.5·v_{max}.

The hyperbola that describes the dependence of the reaction velocity on substrate concentration has three distinct regions (Figure 13.2): for $[S] \ll K_S$, the reaction velocity shows a linear increase with substrate concentration. In this regime, the reaction behaves as a first-order reaction. In this case, [S] in the denominator of eq. 13.9 can be neglected, and we can approximate the reaction velocity as

$$v \approx v_{max} \frac{[S]}{K_S} = k[S] \qquad \text{eq. 13.10}$$

k is the first-order rate constant for the overall reaction. First-order kinetics is observed for $[S] < 0.01 \cdot K_S$. At very high concentrations of [S], the reaction velocity becomes independent of the substrate concentration. In this regime, the reaction follows zero-order kinetics, and $v = v_{max}$. Zero-order kinetics is observed for $[S] > 100 \cdot K_S$. In-between these extreme cases, the reaction velocity exhibits the hyperbolic dependence on substrate concentration expressed by eq. 13.9 (see Section 16.1).

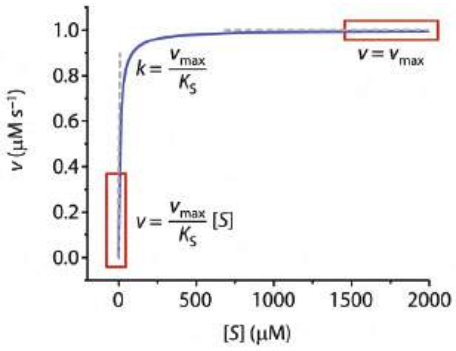


FIGURE 13.2: Zero- and first-order kinetics for high and low substrate concentrations. Reaction velocity v as a function of substrate concentration [S]. For [S] $\ll K_S$, the reaction is first-order, with $k = v_{max}/K_S$. At high concentrations [S] $\gg K_S$, the reaction velocity becomes independent of substrate concentration (zero-order), with $v = v_{max}$. $K_S = 10$ μM, $v_{max} = 1$ μM s^{-1}.

Enzymes are catalysts (see Chapter 16) that catalyze reactions of their substrates in solution. Reactions in which the catalyst is in the same phase of the system as its substrate are termed *homogeneous catalysis*. If the catalyst and the reactant(s) occur in different phases, we speak of *heterogeneous catalysis* (see Chapter 16.1). In both cases, the reaction velocity depends on the fractional saturation of the binding sites available on the catalyst (Section 3.4).

13.2 STEADY-STATE APPROXIMATION (BRIGGS-HALDANE FORMALISM)

The Michaelis-Menten formalism assumes a true thermodynamic equilibrium between enzyme and substrate and the enzyme-substrate complex. This is only the case if k_p is much smaller than the rate constant for its dissociation back to enzyme and substrate (k_{-1} in Scheme 13.3). If $k_p > k_{-1}$, then product formation is rapid once the enzyme-substrate complex has been formed. As we have seen before for consecutive reactions (see Section 8.3), ES will not accumulate in this case, and never reaches the concentration that would correspond to the thermodynamic equilibrium.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_p} E + P \qquad \text{scheme 13.3}$$

According to the principle of Le Chatelier (Section 2.1.2), the withdrawal of ES from the equilibrium by the product-forming reaction shifts the equilibrium to the right, further toward ES. If the substrate is in large excess over enzyme, then, after a short period of time, a *steady state* will be established at which the concentration of ES is constant over time:

$$\frac{d[ES]}{dt} = 0 \qquad \text{eq. 13.11}$$

In this steady-state, ES is produced with the same velocity as it is converted to E and P. According to Scheme 13.3, the change in ES concentration over time is

$$\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_p[ES] \qquad \text{eq. 13.12}$$

By combining eq. 13.12 and eq. 13.11, we obtain

$$k_1[E][S] = (k_{-1} + k_p)[ES] \qquad \text{eq. 13.13}$$

which we can solve for [*ES*]:

$$[ES] = \frac{k_1}{k_{-1} + k_p}[E][S] \qquad \text{eq. 13.14}$$

By defining an apparent (!) equilibrium constant K_M as

$$K_M = \frac{k_{-1} + k_p}{k_1} \qquad \text{eq. 13.15}$$

we can write eq. 13.14 in a similar form as eq. 13.2 for the rapid equilibrium approach:

$$[ES] = \frac{[E][S]}{K_M} \qquad \text{eq. 13.16}$$

The velocity of product formation is

$$v = k_p[ES] \qquad \text{eq. 13.3}$$

We now again divide by the total enzyme concentration $[E]_t$, and express $[E]_t$ as the sum of the concentrations of free enzyme E and enzyme-substrate complex ES (mass conservation). If we now express [*ES*] as a function of the constant K_M (eq. 13.16), we obtain the equivalent expressions to eq. 13.6:

$$\frac{v}{[E]_t} = \frac{k_p \dfrac{[E][S]}{K_M}}{[E] + \dfrac{[E][S]}{K_M}} \qquad \text{eq. 13.17}$$

and

$$\frac{v}{k_p[E]_t} = \frac{\dfrac{[S]}{K_M}}{1 + \dfrac{[S]}{K_M}} \qquad \text{eq. 13.18}$$

which may be simplified to

$$\frac{v}{v_{max}} = \frac{[S]}{K_M + [S]} \qquad \text{eq. 13.19}$$

or

$$v = v_{max} \cdot \frac{[S]}{K_M + [S]} \qquad \text{eq. 13.20}$$

Thus, for the steady-state approach we obtain a velocity equation that describes the hyperbolic dependence of the reaction velocity on the substrate concentration (Figure 13.1) and is formally identical to the rate equation for the rapid equilibrium approach (eq. 13.9). K_M is the substrate concentration at which the reaction occurs with half-maximal velocity. It is important to note that, despite the same shape of the velocity curves, the equilibrium constants K_S and K_M have different meanings. While K_S is the true equilibrium dissociation constant of the ES complex, K_M reflects a ratio of steady-state concentrations. If product formation is rapid ($k_p \gg k_{-1}$), K_M is larger than K_S, and provides an upper limit for K_S. Eq. 13.9 and eq. 13.20 become equivalent for $k_p \ll k_{-1}$, *i.e.* when

product formation becomes rate-limiting, and ES can reach its thermodynamic equilibrium concentration instead of the steady-state equilibrium. In this case, K_M equals K_S, the dissociation constant of the ES complex.

We can determine v_{max}, k_p, and K_M for enzyme-catalyzed reactions by measuring the product concentration as a function of time for different concentrations of substrates, such that $S_0 \gg [E]_t$. The rate of product formation v is then plotted as a function of the substrate concentration $[S] \approx S_0$ (Figure 13.1). The maximum reaction velocity v_{max} is obtained as the saturating value of the resulting hyperbolic curve at high substrate concentrations. From the substrate concentration at which the reaction velocity v is half-maximal, K_M or K_S are determined. The rate constant k_p, also called turnover number k_{cat}, can be calculated according to eq. 13.4 by dividing v_{max} by the total enzyme concentration $[E]_t$.

Historically, eq. 13.20 has been converted into various reciprocal forms such that a linear relationship of some sort between v and S was obtained, and v_{max} and K_M were accessible by linear regression. Examples of these linearized forms are the *Lineweaver-Burk plot*, where the inverse of the reaction velocity, $1/v$, is plotted as a function of $1/[S]$ according to

$$\frac{1}{v} = \frac{1}{v_{max}} + \frac{K_M}{v_{max}[S]} \qquad \text{eq. 13.21}$$

and the *Eadie-Hofstee plot*, where v is plotted as a function of $v/[S]$ according to

$$v = v_{max} - \frac{v \cdot K_M}{[S]} \qquad \text{eq. 13.22}$$

(Box 13.1). We will see in Section 13.6 that the reciprocal forms and the linear plots provide a very simple and intuitive representation of different mechanisms of inhibition. However, conversion of the original data (v as $f([S])$ into reciprocal forms leads to a compression of data points, and alters the associated errors (Figure 13.4). Modern data analysis programs directly describe data points by complex equations without the need for further transformations. The parameters of the fitting equation, here v_{max} and K_M, are varied during the fitting procedure, such that deviations between the equation and the experimental data points are minimized and the data are optimally described. Direct analysis of data avoids inappropriate weighting and error transformations associated with linearization.

BOX 13.1: LINEARIZED FORMS OF THE MICHAELIS-MENTEN EQUATION: LINEWEAVER-BURK AND EADIE-HOFSTEE PLOTS.

In a Lineweaver-Burk plot, $1/v$ is plotted as a function of $1/[S]$. The slope of this plot is K_M/v_{max}, the intercept with the y-axis is $1/v_{max}$ (eq. 13.21). The x-axis is intercepted at $-1/K_M$. For a set of enzyme-catalyzed reactions with the same v_{max}, but different K_M values, the y-axis intercept does not change, but the curves have different slopes and different x-axis intercepts. For a set of reactions with the same K_M, but different v_{max} values, on the other hand, the y-axis intercept and the slope will differ, but the x-axis intercepts are the same (Figure 13.3, Section 13.6).

In an Eadie-Hofstee plot, v is plotted as a function of $v/[S]$. The slope is K_M, the y-axis intercept is v_{max}, and the x-axis intercept is v_{max}/K_M (eq. 13.22). For a set of enzyme-catalyzed reactions with the same v_{max}, but different K_M values, the y-axis intercept does not change, but the curves have different slopes and x-axis intercepts. For a set of reactions with the same K_M, but different v_{max} values, on the other hand, the slope will be the same, but the x- and y-axis intercepts differ (Figure 13.3; Section 13.6).

(*Continued*)

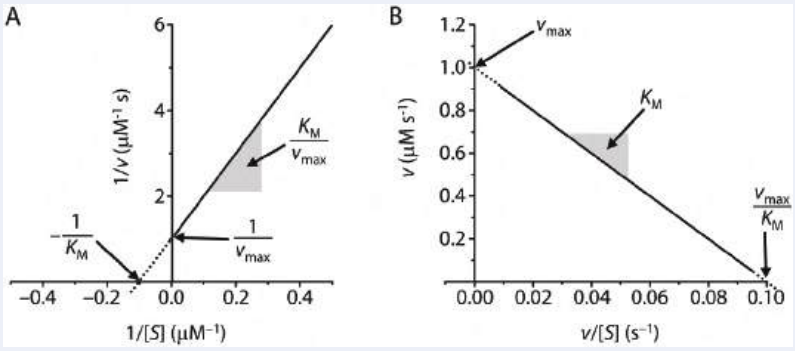


FIGURE 13.3: Reciprocal plots for velocities of enzyme-catalyzed reactions. Lineweaver-Burk (A) and Eadie-Hofstee plots (B) for enzymatic reactions with the same K_M and v_{max} values (K_M = 10 μM, v_{max} = 1 μM s^{-1}). A: The slope of a Lineweaver-Burk plot is K_M/v_{max}, the intercepts with the *y*- and *x*-axes are $1/v_{max}$ and $-1/K_M$. B: The slope of the Eadie-Hofstee plot is K_M, the intercepts with the *y*- and *x*-axes are v_{max} and v_{max}/K_M.

The conversion of measured values into the reciprocal forms leads to a transformation of the associated errors (Figure 13.4). Whenever possible, conversion of data points should therefore be avoided!

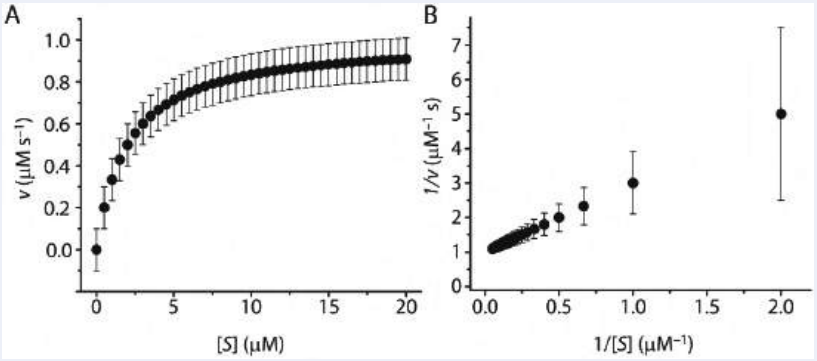


FIGURE 13.4: Error compression in Lineweaver-Burk plots. Direct data with associated errors for the reaction velocities (A) and corresponding Lineweaver-Burk plot (B).

The error propagation from v to $1/v$ is calculated as

$$\sigma_{\frac{1}{v}} = \sqrt{\sigma_v^2 \cdot \left(\frac{\partial \frac{1}{v}}{\partial v}\right)^2} = \sqrt{\sigma_v^2 \cdot \frac{1}{v^4}} \qquad \text{eq. 13.23}$$

(see Section 31.13). The evenly spaced original data points are compressed into the left bottom corner of the plot, and do not contribute much to the definition of the slope. In addition, the propagated errors show that the reciprocal values cannot be weighted identically.

13.3 pH DEPENDENCE

We will discuss the general principles of acid-base catalysis in Section 16.2.1. Here, we consider a reaction where a basic group on the enzyme needs to be protonated to catalyze a reaction. In this case, the reaction velocity will depend on the fraction of the enzyme that is protonated. The reaction will occur at maximum velocity at pH values where the group is fully protonated, *i.e.* at pH values at least one unit below the pK_A value of the basic group (Section 2.3). At higher pH values, above the pK_A value, the basic group on the enzyme will become deprotonated, which depletes the active species and reduces the reaction velocity (Figure 13.5).

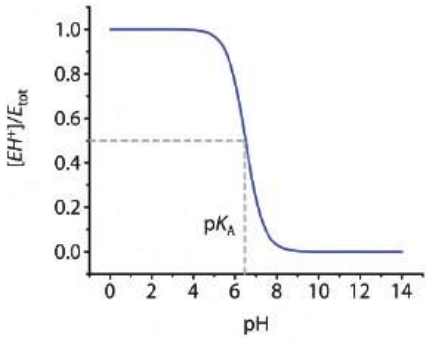
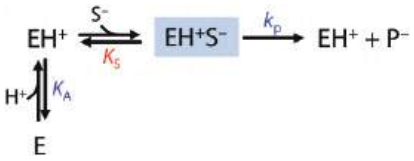


FIGURE 13.5: Protonated and deprotonated states of an enzyme as a function of pH. The fraction of protonated enzyme as a function of pH is depicted for a pK_A value of 6.5. At pH values well below the pK_A, the enzyme is in the protonated active state, and will show maximal activity. At pH values well above the pK_A, the enzyme is in the deprotonated, inactive state and will show no activity.

We now derive the rate equation for an enzymatic reaction where EH^+ is the active form that binds to a negatively charged substrate S^-. Product formation occurs from the EH^+S^- complex with the rate constant k_p (Scheme 13.4):



scheme 13.4

We will follow a widely applicable procedure to derive the rate equation, the reaction velocity as a function of substrate concentration and apparent v_{max} and K_M values, for reactions as a function of pH. The standard sequence of steps to derive these equations is outlined in general terms in Box 13.2. We will apply this procedure to enzymes with multiple substrate binding sites (Sections 13.4, 13.5), and to enzyme-catalyzed reactions in the presence of inhibitors (Section 13.6).

The rate of product formation by the enzyme depends on the rate constant k_p for product formation and the concentration of the active enzyme complex $[EH^+S^-]$ that catalyzes product formation. The reaction velocity v is thus

$$v = k_p\left[EH^+S^-\right] \qquad \text{eq. 13.24}$$

We can now divide by the total enzyme concentration $[E]_t$ on both sides of the equation. On the left-hand side, we do not further specify $[E]_t$ but on the right-hand side we write $[E]_t$ as the sum of the concentrations of all possible complexes the enzyme can form (eq. 13.25). Again, this expression relates the macroscopic property, the reaction velocity v, to the microscopic rate constant, k_p, and the fraction of enzyme molecules in the catalytically active state, $[EH^+S^-]$, relative to all possible states (see Section 3.4):

$$\frac{v}{[E]_t} = \frac{k_p[EH^+S^-]}{[E]+[EH^+]+[EH^+S^-]} \qquad \text{eq. 13.25}$$

We now divide both sides by k_p, and obtain

$$\frac{v}{k_p[E]_t} = \frac{v}{v_{max}} = \frac{[EH^+S^-]}{[E]+[EH^+]+[EH^+S^-]} \qquad \text{eq. 13.26}$$

The product $k_p \cdot [E]_t$ corresponds to the maximum velocity of the reaction, v_{max}, under conditions when all enzyme is protonated and saturated with substrate. As pointed out before (Section 13.1), eq. 13.26 relates the relative velocity, v/v_{max}, to the fraction of substrate-bound enzyme in the active form, in this case the protonated state. In the next step, we express all concentrations of enzyme complexes on the right-hand side in terms of the equilibrium dissociation constants K_A and K_S. The protonated enzyme concentration $[EH^+]$ is

$$[EH^+] = \frac{[E][H^+]}{K_A} \qquad \text{eq. 13.27}$$

and the concentration of the protonated and substrate-bound enzyme $[EH^+S^-]$ is

$$[EH^+S^-] = \frac{[EH^+][S^-]}{K_S} = \frac{[H^+][E][S^-]}{K_A K_S} \qquad \text{eq. 13.28}$$

Substitution of these expressions into eq. 13.26 gives

$$\frac{v}{v_{max}} = \frac{\frac{[E][H^+][S^-]}{K_A K_S}}{[E]+\frac{[E][H^+]}{K_A}+\frac{[E][H^+][S^-]}{K_A K_S}} \qquad \text{eq. 13.29}$$

By canceling $[E]$ on the right-hand side, this expression can be simplified to

$$\frac{v}{v_{max}} = \frac{\frac{[H^+][S^-]}{K_A K_S}}{1+\frac{[H^+]}{K_A}+\frac{[H^+][S^-]}{K_A K_S}} \qquad \text{eq. 13.30}$$

Again, the right-hand side is the probability to find the system in the active, protonated, and substrate-bound state; the denominator is the partition function of the system, with the enzyme E as the reference state (see Sections 3.1 and 3.2). The proton concentration $[H^+]$ on the right-hand side of eq. 13.30 can be canceled, and eq. 13.30 can be rearranged to

$$\frac{v}{v_{max}} = \frac{[S^-]}{\frac{K_A K_S}{[H^+]}+K_S+[S^-]} = \frac{[S^-]}{K_S\left(1+\frac{K_A}{[H^+]}\right)+[S^-]} \qquad \text{eq. 13.31}$$

Eq. 13.31 can be expressed as

$$\frac{v}{v_{max}} = \frac{[S^-]}{K_{M,app}+[S^-]} \qquad \text{eq. 13.32}$$

with

$$K_{M,app} = K_S\left(1 + \frac{K_A}{[H^+]}\right) \qquad \text{eq. 13.33}$$

From eq. 13.32 and eq. 13.33, we see that v_{max} is not pH-dependent, but $K_{M,app}$ now is an apparent Michaelis-Menten constant that depends on the K_A value of the basic group and on the proton concentration, *i.e.* the pH. At high concentrations of protons (low pH), $K_{M,app}$ approaches K_S, because all enzyme is in the protonated state and can bind the substrate S^-. At low concentrations of protons, (high pH), $K_{M,app}$ becomes larger than K_S. The increase in $K_{M,app}$ reflects the decrease in substrate binding when only a small fraction of the enzyme is protonated.

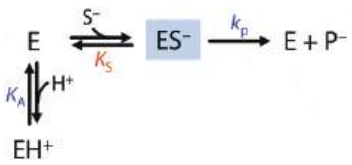
BOX 13.2: A STANDARD PROCEDURE TO DERIVE VELOCITY EQUATIONS AND APPARENT K_M AND v_{max} VALUES.

For more complex reaction schemes that comprise more reactions than just substrate binding and product formation, expressions for v/v_{max} as a function of substrate concentration can be obtained by a standard procedure following these steps:

1. Description of the reaction velocity v as a function of the concentration of active complexes of the enzyme that catalyze product formation (eq. 13.24). If more than one complex catalyzes product formation, the concentrations of these complexes have to be added.
2. Division of both sides of the equation by the total enzyme concentration $[E]_t$. On the right-hand side, $[E]_t$ is expressed as the sum of the concentrations of all complexes the enzyme can form (eq. 13.25).
3. Division by k_p gives v/v_{max} on the left-hand side (eq. 13.26).
4. Expression of all concentrations of enzyme complexes in terms of $[E]$, $[S]$, and dissociation constants (eq. 13.27-eq. 13.29) and canceling $[E]$ from all terms (eq. 13.30).
5. Simplification and rearrangement to obtain an equation in the form of the Michaelis-Menten equation (eq. 13.31, eq. 13.32).
6. Relating $v_{max,app}$ to v_{max}: division of the equation by the factor before $[S]$ in the numerator to obtain an expression for $v_{max,app}$ (not applicable if $v_{max,app} = v_{max}$).
7. Relating $K_{M,app}$ to K_S: summary of all terms added to $[S]$ in the denominator to obtain an expression for $K_{M,app}$ (eq. 13.33).

This procedure can be used to derive expressions for the reaction velocity as a function of substrate concentration and for apparent v_{max} and K_M values for enzyme-catalyzed reactions as a function of pH, in the presence of inhibitors (Section 13.6) or activators, and for enzymes with multiple substrate binding sites (Sections 13.4, 13.5).

We can derive the corresponding form of eq. 13.33 for a scenario where the deprotonated enzyme, but not the protonated enzyme, binds substrate (Scheme 13.5).



scheme 13.5

In this case, we can derive $v_{max,app} = v_{max}$, and

$$K_{M,app} = K_S\left(1+\frac{[H^+]}{K_A}\right) \qquad \text{eq. 13.34}$$

Here, $K_{M,app}$ becomes equal to K_S for high pH (low $[H^+]$), where the enzyme is mostly deprotonated and binds substrate. For low pH (high $[H^+]$), where most of the enzyme is protonated and does not bind substrate, $K_{M,app}$ is larger than K_S (Figure 13.6). Formally, this behavior is identical to competitive inhibition, with protons acting as a competitive inhibitor (see Section 13.6.2).

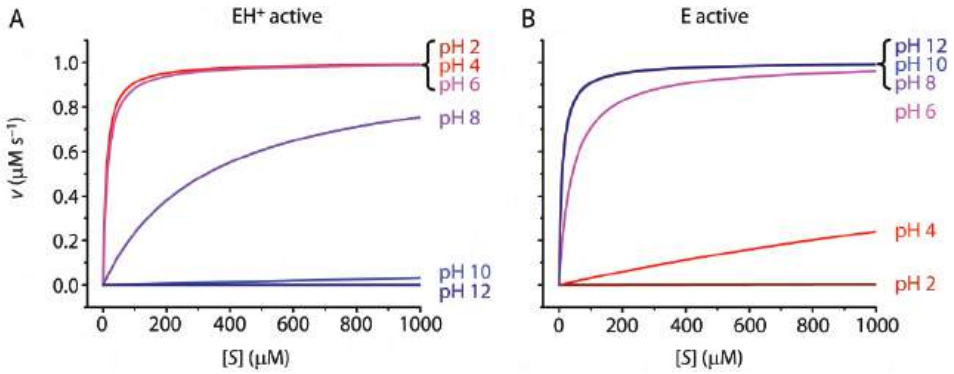


FIGURE 13.6: Influence of protolysis equilibria on $K_{M,app}$. Reaction velocity v as a function of substrate concentration at different pH. A: The protonated state EH^+ is the active species (Scheme 13.4). At low pH, the enzyme is protonated and active. With increasing pH, the enzyme becomes deprotonated, and the $K_{M,app}$ value increases (eq. 13.33). B: The deprotonated enzyme E is the active species (Scheme 13.5). At high pH, the enzyme is deprotonated and active. With decreasing pH, the enzyme becomes protonated, and $K_{M,app}$ increases (eq. 13.34). $pK_A = 6.5$ ($K_A = 3.6 \cdot 10^{-7}$ M).

It is also possible that both enzyme and substrate need to be in the correct protonated/deprotonated state for the reaction to occur, *e.g.* if the substrate is a weak acid that is only bound in its deprotonated S^- form, and the enzyme carries a basic group in its active site that needs to be protonated for the substrate to bind. The maximum activity would then be observed at pH values where all the enzyme exists in the protonated state as EH^+ and all the substrate exists in the deprotonated state as S^-. This condition may not be possible if the two pK_A values of the groups involved are too similar (Figure 13.7).

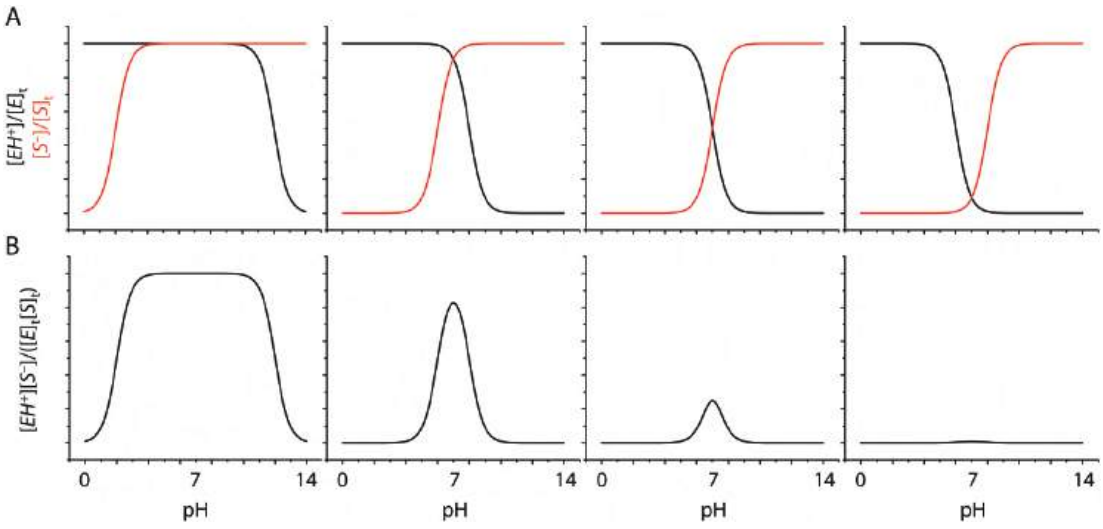
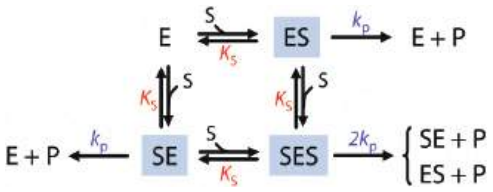


FIGURE 13.7: Protonated and deprotonated states of enzyme and substrate and reaction velocity as a function of pH. A: Fraction protonated enzyme (EH^+, black) and deprotonated substrate (S^-, red) as a function of the pH. pK_A values: 12 (enzyme)/2 (substrate), 8/6, 7/7, and 6/8. B: The enzyme is active in its protonated state, the substrate has to be deprotonated to be bound. The product of the fractions of protonated enzyme and deprotonated enzyme as a function of pH (black) equals the probability of an encounter between EH^+ and S^-, and is thus a measure of the reaction velocity.

13.4 TWO OR MORE NON-INTERACTING ACTIVE SITES

So far, we have limited our considerations to enzymes with one active site. Often, enzymes assemble into dimeric, trimeric, tetrameric, or higher oligomeric states, combining more than one active site in one entity. We first look at a dimeric enzyme that contains two active sites (Scheme 13.6). For simplicity, we now refer to the dimeric enzyme as E, with E binding one or two substrate molecules S. Thus, the enzyme can form an ES, an SE, or an SES complex. We assume first that the active sites are independent, meaning the dissociation of substrates from each site is described by K_S, and they all catalyze conversion to products with k_p.



scheme 13.6

When we express the reaction velocity in terms of rate constants and concentrations of relevant species, we now have to take into account all enzyme complexes from which product is formed, and the number of active sites leading to product formation in the respective complex. For Scheme 13.6, the complexes ES, SE, and SES catalyze product formation. While ES and SE convert one substrate into the respective product, the SES complex turns over two substrate molecules to two molecules of the product, and generates product twice as rapid as the ES and SE complexes. The overall reaction velocity is thus described by eq. 13.35:

$$v = k_p[ES] + k_p[SE] + 2k_p[SES] \qquad \text{eq. 13.35}$$

Again, we divide by the total enzyme concentration $[E]_t$, specify the individual complexes contributing to $[E]_t$ on the right-hand side of the equation only, and obtain

$$\frac{v}{[E]_t} = \frac{k_p[ES] + k_p[SE] + 2k_p[SES]}{[E] + [ES] + [SE] + [SES]} \qquad \text{eq. 13.36}$$

The enzyme complexes are expressed as

$$[ES] = [SE] = \frac{[S][E]}{K_S} \qquad \text{eq. 13.37}$$

and

$$[SES] = \frac{[S][ES]}{K_S} = \frac{[S]^2}{K_S^2}[E] \qquad \text{eq. 13.38}$$

according to the definition of the equilibrium dissociation constant K_S. Substitution into eq. 13.36, division of all terms on the right-hand side by the enzyme concentration $[E]$, and rearrangement yields

$$\frac{v}{[E]_t} = \frac{k_p\frac{[S]}{K_S} + k_p\frac{[S]}{K_S} + 2k_p\frac{[S]^2}{K_S^2}}{1 + \frac{[S]}{K_S} + \frac{[S]}{K_S} + \frac{[S]^2}{K_S^2}} = \frac{2 \cdot k_p\frac{[S]}{K_S} + 2k_p\frac{[S]^2}{K_S^2}}{1 + 2\frac{[S]}{K_S} + \frac{[S]^2}{K_S^2}} \qquad \text{eq. 13.39}$$

and

$$\frac{v}{2k_p[E]_t} = \frac{v}{v_{max}} = \frac{\frac{[S]}{K_S} + \frac{[S]^2}{K_S^2}}{1 + 2\frac{[S]}{K_S} + \frac{[S]^2}{K_S^2}} \qquad \text{eq. 13.40}$$

with

$$2k_p[E]_t = v_{max} \qquad \text{eq. 13.41}$$

The maximum velocity v_{max} is now twice as large as for the same enzyme with one active site, because both active sites in the dimer contribute to product formation. Again, the right-hand side of eq. 13.40 represents the probability for an occupied binding site; the denominator corresponds to the partition function for the system, with the free enzyme E as the reference state, now taking into account the degeneracy of the state with one of the two binding sites occupied (Sections 3.1 and 3.2).

If we look back at where the individual terms in eq. 13.39 come from in the derivation, we see that the two $[S]/K_S$ terms in the numerator reflect the two species that catalyze product formation, ES and SES. The exponent of each $[S]/K_S$ term corresponds to the number of binding sites occupied (1 for ES_1, 2 for ES_2). The coefficients of each term contain two components: the rate constant of product formation by the respective complex, and the number of possibilities to generate this species. For two active sites, there are two ways to generate an ES complex (ES and SE, in the following ES_1), but only one possibility to generate an ES_2 complex. The rate constant of product formation for ES_1 is k_p. The rate constant for ES_2 is $2k_p$ because both active sites generate product. Finally, the three terms in the denominator represent all forms of the enzyme, free enzyme (E), enzyme with one substrate molecule bound (ES_1), and enzyme with two substrate molecules bound (ES_2). Each enzyme-substrate complex is reflected by an $[S]/K_S$ term with the exponent corresponding to the number of binding sites occupied. Note that x^0 equals unity, thus the free enzyme is reflected by $([S]/K_S)^0 = 1$, it is an enzyme-substrate complex with no substrate bound, so to say. The coefficients in front of the $[S]/K_S$ terms correspond to the number of possibilities in which the species can be generated, *i.e.* the degeneracy of this microstate (see Section 3.2.2), which is two for ES_1, and one for ES_2 (Figure 13.8).

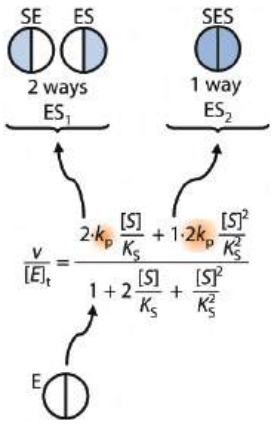


FIGURE 13.8: Enzyme-substrate complexes for an enzyme with two substrate binding sites. In the numerator, each $[S]/K_S$ term represents a substrate-bound enzyme species. The number of occupied substrate binding sites (highlighted in blue) is reflected in the exponent of the $[S]/K_S$ term. Each term is multiplied by the number of different ways to generate the respective complex (2 ways for ES_1, one way for ES_2) and by the rate constant of product formation for each complex (k_p for ES_1, $2k_p$ for ES_2, orange). The denominator contains one $[S]/K_S$ term for each species of enzyme present (E, ES_1, ES_2). Again, the exponent specifies the number of binding sites occupied, giving 1 for the free enzyme ($x^0 = 1$). Each term is multiplied by the number of ways to generate the respective complex (1, 2, and 1 for E, ES_1, and ES_2). Division by $2k_p$, the rate constant for the ES_2 complex that is fully occupied with substrate, then gives v_{max} in the denominator on the left-hand side of the equation.

These coefficients are the binomial coefficients that are found in Pascal's triangle (see Section 31.3). The fraction on the right-hand side of eq. 13.40 thus represents the number of binding sites occupied, divided by the total number of binding sites present, which is nothing else than the fractional saturation θ of the binding sites available on the enzyme (Sections 3.4 and 13.1).

Recognizing this principle, we can now directly construct the equation for an enzyme with four non-interacting catalytic sites. On the left-hand side we write $v/[E]_t$. On the right-hand side, we have to consider all species that catalyze product formation in the numerator. For an enzyme with four active sites, we have enzyme with one, two, three, or four substrates bound (ES, ES_2, ES_3, ES_4, respectively), represented by $([S]/K_S)$, $([S]/K_S)^2$, $([S]/K_S)^3$, and $([S]/K_S)^4$. The factors for each term depend on the number of possibilities to generate the respective complex. There is only one way to populate an ES or ES_4 complex, but there are six different possibilities for ES_2 and four for ES_3. With these considerations, we have completed the numerator. For the denominator, we have to take into account free enzyme E, ES, ES_2, ES_3, and ES_4 represented by $([S]/K_S)^n$ with n = 0, 1, 2, 3, and 4. The coefficients are the numbers of ways to generate the respective species, 1 for E, 4 for ES_1, 6 for ES_2, 4 for ES_3, and 1 for ES_4 (Figure 13.9): the binomial coefficients that are found in Pascal's triangle (see Section 31.3).

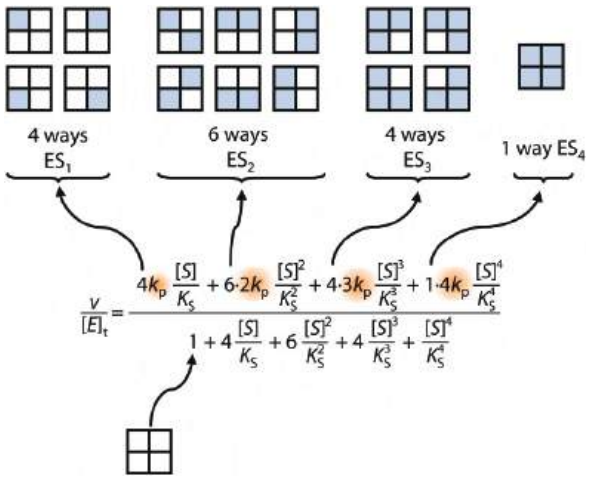


FIGURE 13.9: Enzyme-substrate complexes for an enzyme with four substrate binding sites. In the numerator, each $[S]/K_S$ term represents a substrate-bound enzyme species (ES_1, ES_2, ES_3, ES_4, occupied substrate binding sites highlighted in blue). The exponent of the $[S]/K_S$ term is the respective number of occupied substrate binding sites. Each term is multiplied by the number of different ways to generate the respective complex (4 ways for ES_1, 6 ways for ES_2, 4 ways for ES_3, 1 way for ES_4), and by the rate constant of product formation for each complex (k_p, $2k_p$, $3k_p$, and $4k_p$ (orange) for ES_1, ES_2, ES_3, ES_4). The denominator contains one $[S]/K_S$ term for each species of enzyme present (E, ES_1, ES_2, ES_3, ES_4), with the exponent equal to the number of binding sites occupied. Each term is multiplied by the number of ways to generate the respective complex (1, 4, 6, 4, 1 for E, ES_1, ES_2, ES_3, ES_4). Division by $4k_p$, the rate constant for product formation from the ES_4 complex that is fully occupied with substrate, then gives v_{max} in the denominator on the left-hand side of the equation.

Altogether, we obtain eq. 13.42,

$$\frac{v}{[E]_t} = \frac{4k_p\frac{[S]}{K_S} + 6\cdot 2k_p\frac{[S]^2}{K_S^2} + 4\cdot 3k_p\frac{[S]^3}{K_S^3} + 1\cdot 4k_p\frac{[S]^4}{K_S^4}}{1+4\frac{[S]}{K_S}+6\frac{[S]^2}{K_S^2}+4\frac{[S]^3}{K_S^3}+\frac{[S]^4}{K_S^4}} \qquad \text{eq. 13.42}$$

Dividing eq. 13.42 by $4k_p$ leaves us with

$$v_{max} = 4k_p[E]_t \qquad \text{eq. 13.43}$$

in the denominator on the left, and smaller coefficients in the numerator on the right:

$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S}+3\frac{[S]^2}{K_S^2}+3\frac{[S]^3}{K_S^3}+\frac{[S]^4}{K_S^4}}{1+4\frac{[S]}{K_S}+6\frac{[S]^2}{K_S^2}+4\frac{[S]^3}{K_S^3}+\frac{[S]^4}{K_S^4}} \qquad \text{eq. 13.44}$$

We can write the sums in the numerator and denominator of eq. 13.44 as binomials (eq. 13.45)

$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S}\left(1+\frac{[S]}{K_S}\right)^3}{\left(1+\frac{[S]}{K_S}\right)^4} \qquad \text{eq. 13.45}$$

for four active sites. Eq. 13.45 can be generalized to n active sites:

$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S}\left(1+\frac{[S]}{K_S}\right)^{n-1}}{\left(1+\frac{[S]}{K_S}\right)^n} \qquad \text{eq. 13.46}$$

By eliminating $(1+[S]/K_S)^{n-1}$, eq. 13.46 reduces to eq. 13.47:

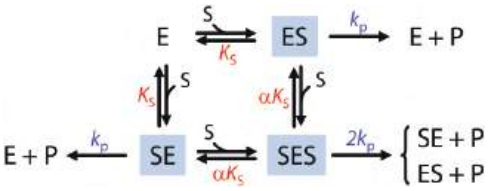
$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S}}{1+\frac{[S]}{K_S}} = \frac{[S]}{K_S+[S]} \qquad \text{eq. 13.47}$$

which is identical to eq. 13.19 for an enzyme with a single active site. Thus, an enzyme with n independent active sites shows the same hyperbolic dependence of the reaction velocity on the substrate concentration as an enzyme with only one active site (see Section 3.4.1). We therefore cannot derive the number of active sites on the enzyme from a measurement of v as a function of $[S]$. To determine the substrate binding stoichiometry, we have to resort to equilibrium titrations (active-site titrations, see Section 21.5.5.1).

Overall, the inductive approach used here to derive the reaction velocity as a function of substrate concentration for two, four, and n non-interacting substrate binding sites is equivalent to the analytical treatment of multiple binding equilibria, based on the concepts of statistical thermodynamics (Section 3.4.1).

13.5 TWO OR MORE INTERACTING ACTIVE SITES: COOPERATIVITY AND THE HILL EQUATION

The rate equations for enzymes with one or more active sites become different when the individual active sites of the enzyme interact with each other. The interaction between two active sites can be positive or negative: the binding of substrate to one active site can promote or inhibit binding of a second substrate molecule to the second active site. If the equilibrium dissociation constant for the ES_1 complex is K_S, we can describe the second equilibrium constant for dissociation of one substrate from ES_2 ($ES_2 \rightarrow ES_1 + S$) as $\alpha \cdot K_S$ (Scheme 13.7). The parameter α takes into account the interaction between the two active sites, and is a measure of the *cooperativity* of the enzyme. A positive interaction or *positive cooperativity* is expressed by $\alpha < 1$ (corresponding to a smaller dissociation constant K_S, or larger binding constant, for substrate 2; Sections 2.2, 2.4): binding of one substrate molecule facilitates binding of the second substrate by α-fold. A negative interaction is expressed by $\alpha > 1$ (corresponding to a larger dissociation constant K_S, or smaller binding constant, for substrate 2; *negative cooperativity*). In this case, binding of the first substrate leads to an α-fold reduction in affinity for the second substrate molecule (see Section 3.4.3).



scheme 13.7

In the equation for the reaction velocity, we have to take into account the factor α for the second binding event, *i.e.* in all terms that represent the SES complex, which leads to eq. 13.48:

$$\frac{v}{[E]_t} = \frac{2k_p \frac{[S]}{K_S} + 2k_p \frac{[S]^2}{\alpha K_S^2}}{1 + 2\frac{[S]}{K_S} + \frac{[S]^2}{\alpha K_S^2}} \qquad \text{eq. 13.48}$$

or

$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S} + \frac{[S]^2}{\alpha K_S^2}}{1 + 2\frac{[S]}{K_S} + \frac{[S]^2}{\alpha K_S^2}} \qquad \text{eq. 13.49}$$

As in eq. 13.40 for an enzyme with two non-interacting sites, the fraction on the right-hand side of eq. 13.49 represents the fractional saturation θ of the enzyme with substrate, this time taking into account the cooperativity between the individual sites. However, in eq. 13.49, the sums in the numerator and denominator cannot be expressed as binomials, and the equation therefore does not reduce to a simple hyperbolic form! Thus, the dependence of the reaction velocity v on the substrate concentration $[S]$ will be different if there is cooperativity between the active sites (Figure 13.10; see Section 3.4.3). According to eq. 13.49, the reaction velocity v shows a sigmoidal dependence on the substrate concentration. While the end value at substrate saturation is still v_{max}, K_S now does not represent the concentration at which the half-maximal velocity is reached.

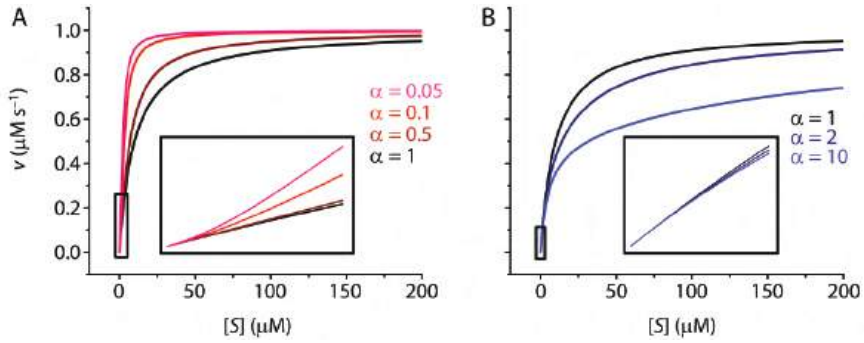


FIGURE 13.10: Cooperativity and sigmoidal dependence of the reaction velocity on substrate concentration. Reaction velocity *v* as a function of substrate concentration [*S*] for an enzyme with two substrate binding sites and cooperativity. A: Positive cooperativity. The cooperativity parameter α is 1 (no cooperativity, black), 0.5 (brown), 0.1 (red), and 0.05 (pink). The inset shows the region for small [*S*] where the differences are most pronounced: with increasing cooperativity (decreasing cooperativity parameter α), the curves become more and more sigmoidal, and reach v_{max} at lower substrate concentration. B: Negative cooperativity. The cooperativity parameter α is 1 (no cooperativity, black), 2 (dark blue), and 10 (blue). The inset shows the region for small [*S*]: all curves show a hyperbolic dependence on substrate concentration.

Cooperativity between multiple binding sites can be described by concerted and sequential models (Section 3.4.3). In the *sequential model* (Koshland-Nemethy-Filmer model, Section 3.4.3.2), binding of a substrate molecule only alters the affinity of this site, and mixed species with high- and low-affinity binding sites are populated (Scheme 13.8; see Figure 13.11).

To describe the binding of four ligand molecules to four cooperative sites following the sequential model, we have to introduce a factor to take into account the interaction of the sites for each binding step, *i.e.* binding of the first substrate molecule favors/disfavors binding of the second by α-fold, binding of the second in turn favors/disfavors binding of the third ligand by β-fold, and binding of the third ligand favors/disfavors binding of the fourth substrate molecule by γ-fold (Scheme 13.8).



scheme 13.8

In this case, eq. 13.49 becomes

$$\frac{v}{v_{max}} = \frac{\frac{[S]}{K_S} + 3\frac{[S]^2}{\alpha K_S^2} + 3\frac{[S]^3}{\alpha^2\beta K_S^3} + \frac{[S]^4}{\alpha^3\beta^2\gamma K_S^4}}{1 + 4\frac{[S]}{K_S} + 6\frac{[S]^2}{\alpha K_S^2} + 4\frac{[S]^3}{\alpha^2\beta K_S^3} + \frac{[S]^4}{\alpha^3\beta^2\gamma K_S^4}} \qquad \text{eq. 13.50}$$

and the dependence of *v* on [*S*] will again be sigmoidal (Figure 13.11). An example for cooperative binding to four binding sites is hemoglobin, which contains four interacting binding sites for its substrate oxygen.

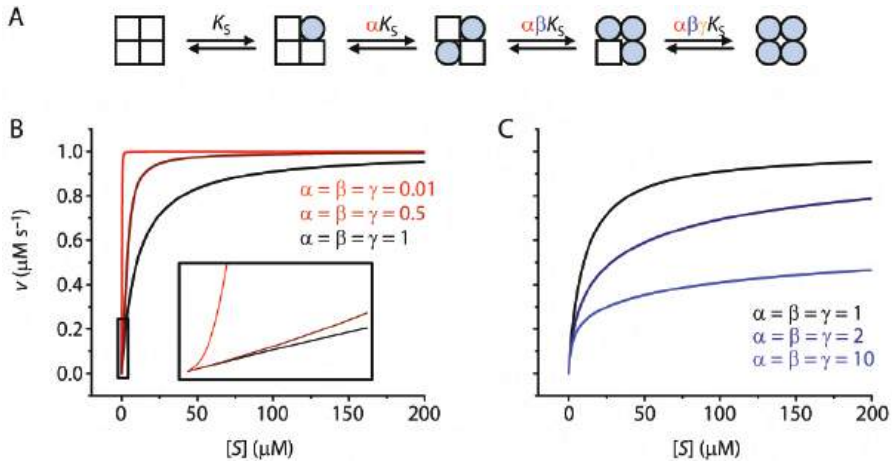


FIGURE 13.11: Sigmoidal dependence of reaction velocities on substrate concentration for four cooperative sites. A: Binding scheme. B: Reaction velocity v as a function of substrate concentration [S] for an enzyme with four substrate binding sites and cooperativity parameters α, β, and γ of 1 (no cooperativity, black), 0.5 (brown), and 0.01 (red). The inset shows the region for small [S] where the differences are most pronounced: with increasing cooperativity (decreasing cooperativity parameters), the curves become more and more sigmoidal, and reach v_{max} at lower substrate concentration. C: Reaction velocity v as a function of substrate concentration [S] for an enzyme with four substrate binding sites and negative cooperativity: α, β, and γ = 1 (no cooperativity, black), 2 (dark blue), and 10 (blue). All curves are hyperbolic.

In contrast to the sequential model, cooperativity can also be rationalized in terms of a *concerted model* (Monod-Wyman-Changeux model, see Section 3.4.3.1). In the concerted model, binding of the first ligand converts all subunits of the enzyme into a high-affinity state, leading to the same sigmoidal dependence of the reaction velocity on substrate concentration as we derived for the sequential model. The concerted model explains positive cooperativity, but cannot describe negative cooperativity. A general treatment of cooperative binding and the MWC and KNF models, based on the considerations of statistical thermodynamics, is provided in Section 3.4.

Let us stay with an enzyme with four active sites, but now we consider a case where the cooperativity is so high that binding of one substrate molecule will immediately lead to saturation of the three remaining binding sites with substrate. In this case, the enzyme only exists in the free form or in the form saturated with substrate molecules, which will simplify things tremendously. The intermediate states with one, two, or three substrate molecules bound will not be significantly populated. Thus, we do not have to consider them in eq. 13.50, and can simplify this expression to

$$\frac{v}{v_{max}} = \frac{\frac{[S]^4}{\alpha^3\beta^2\gamma K_S^4}}{1+\frac{[S]^4}{\alpha^3\beta^2\gamma K_S^4}} = \frac{\frac{[S]^4}{K'}}{1+\frac{[S]^4}{K'}} = \frac{[S]^4}{K'+[S]^4} \qquad \text{eq. 13.51}$$

with

$$v_{max} = 4k_p[E]_t \qquad \text{eq. 13.43}$$

and

$$K' = \alpha^3\beta^2\gamma K_S^4 \qquad \text{eq. 13.52}$$

Eq. 13.51 is called the *Hill equation*. The Hill equation can be generalized to n interacting sites as

$$\frac{v}{v_{max}} = \frac{[S]^n}{K' + [S]^n} \qquad \text{eq. 13.53}$$

with

$$K' = K_S^n \left(\alpha^{n-1}\beta^{n-2}\gamma^{n-3}\ldots\right) \qquad \text{eq. 13.54}$$

The parameter n is called the *Hill coefficient*. For the limiting case of high cooperativity, n equals the number of binding sites. If we have intermediate or low cooperativity, we can still use the Hill equation to describe the dependence of the reaction velocity on the substrate concentration (Figure 13.12), and can extract a value for the Hill coefficient n. However, in these cases n does not correspond to the number of binding sites but merely provides a lower limit for the number of binding sites present (Box 13.3). The Hill equation does not describe negative cooperativity. Hill parameters below unity hence have no physical meaning.

Note that the Hill equation should never be used to describe the dependence of the reaction velocity on the substrate concentration for enzymes containing a single substrate binding site! With a single binding site, the limiting conditions for which we derived the Hill equation, *i.e.* the presence of multiple highly cooperative binding sites does not apply. In other words, the Hill equation is based on a model that is not physically meaningful for the case of a single binding site. If we nevertheless use the Hill equation to describe our data, the Hill parameter n just becomes an additional degree of freedom, which may lead to (over)fitted curves that naturally describe the data better than a hyperbola. However, the Hill parameter determined in this analysis does not have any meaning. It essentially acts as a fudge factor.

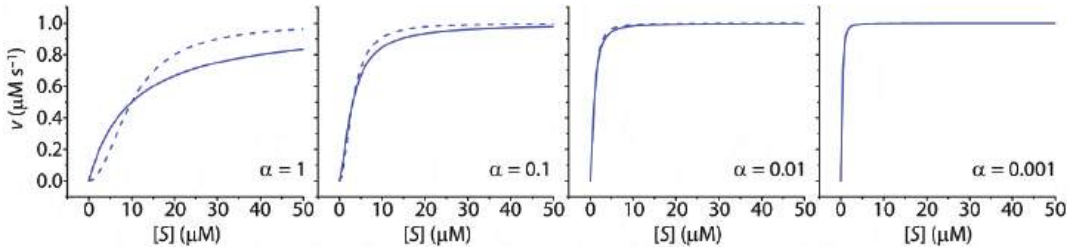


FIGURE 13.12: Approximation by the Hill equation for different levels of cooperativity. The reaction velocity of an enzyme with two substrate binding sites as a function of the substrate concentration [*S*] is calculated explicitly (eq. 13.50, blue), or according to the Hill approximation (eq. 13.53 with $n = 2$, blue dashed line). The cooperativity parameter α is 1 (no cooperativity), 0.1, 0.01, and 0.001. The Hill equation is a good approximation for high cooperativity (α = 0.01, α = 0.001).

BOX 13.3: THE HILL COEFFICIENT AND THE NUMBER OF BINDING SITES.

For enzymes with multiple binding sites and low to intermediate cooperativity, the dependence of the reaction velocity on the substrate concentration can still be described satisfactorily by the Hill equation. The Hill coefficient retrieved by curve-fitting to experimental data will not be an integer value in these cases, however, and is often called the *apparent Hill coefficient* n_{app}. For an enzyme with two low to intermediately cooperative sites (α = 0.5, 0.2, 0.1, 0.01), the extracted Hill coefficients n_{app} are 1.17, 1.37, 1.49, 1.78 (Figure 13.13). For high cooperativity (α = 0.001), a Hill coefficient of 1.97, *i.e.* close to the number of binding sites, is extracted, whereas in the absence of cooperativity (α = 1), the Hill coefficient reduces to $n_{app} = 1$.

(*Continued*)

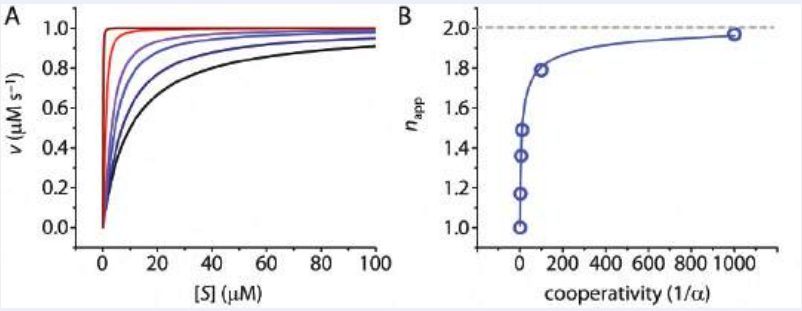


FIGURE 13.13: The apparent Hill coefficient. A: Reaction velocity as a function of substrate concentration for an enzyme with two cooperative sites with cooperativity parameter α = 1, 0.5, 0.2, 0.1, 0.01, 0.001 (from black over blue and red to brown). Non-linear least-square fitting with the Hill equation (eq. 13.53) gives n_{app} = 1, 1.17, 1.37, 1.49, 1.78, and 1.97. B: Apparent Hill coefficient n_{app} as a function of the cooperativity (1/α). For high cooperativity, n_{app} approaches the number of binding sites.

Even for low to intermediate cooperativity, we can still obtain some information from the apparent Hill coefficient n_{app}: the next larger integer $n > n_{app}$ is the minimum number of binding sites. A Hill coefficient of 1.5 can indicate that we are looking at an enzyme with two binding sites that show appreciable cooperativity, at an enzyme with four binding sites and moderate to low cooperativity, or at an enzyme with numerous binding sites and low cooperativity.

Hemoglobin is a tetrameric protein that contains four binding sites for oxygen. It functions in oxygen transport in the blood stream. The four oxygen binding sites of the hemoglobin tetramer show intermediate cooperativity. From the oxygen binding curve, an apparent Hill coefficient of n_{app} = 2.8 can be determined.

13.6 INHIBITION OF ENZYME ACTIVITY

Enzyme activity is often regulated by small molecules that bind either to the same binding site as the substrate or to a different site. Product inhibition (Section 13.6.1) is a common type of competitive inhibition, where the inhibitor interacts reversibly with the active site (Section 13.6.2). Non-competitive inhibitors interact reversibly with a site away from the substrate binding site (Section 13.6.3). Mixed forms are also possible (Section 13.6.4). The enzyme-inhibitor interaction can be quantified through equilibrium constants (K_I) or as IC_{50} values (Section 13.6.5), concentrations at which the reaction velocity is 50% of the maximum velocity v_{max}. Inhibitors can also interact with enzymes slowly, essentially irreversibly, leading to time-dependent inhibition (Section 13.6.6).

13.6.1 PRODUCT INHIBITION IN REVERSIBLE REACTIONS

Many enzymes are inhibited by the products of the reaction they catalyze, leading to a negative feedback on enzyme activity. Product inhibition is a common mechanism for the regulation of enzyme activity. We assume that enzyme and substrate on the one side and enzyme and product on the other side are in rapid equilibrium, characterized by the dissociation constants of ES and EP, K_S and K_P. Forward (product formation) and reverse reactions occur with rate constants k_f and k_r (Scheme 13.9).

$$E + S \underset{K_S}{\rightleftharpoons} ES \underset{k_r}{\overset{k_f}{\rightleftharpoons}} EP \overset{K_P}{\rightleftharpoons} E + P \qquad \text{scheme 13.9}$$

When we consider initial velocities only, *i.e.* we focus on the initial part of a reaction starting from substrate S, where no product has been formed yet ($[P](0) = 0$), the forward reaction occurs at the velocity v_f

$$v_f = v_{max,f} \frac{[S]}{K_S + [S]} \qquad \text{eq. 13.55}$$

If we start the reaction from the right-hand side by supplying the enzyme with product P instead of substrate S, the initial reaction velocity v_r for the reverse reaction is

$$v_r = v_{max,r} \frac{[P]}{K_P + [P]} \qquad \text{eq. 13.56}$$

Enzyme and substrate on the one side and enzyme and product on the other side of the reaction are in rapid equilibrium, and k_f and k_r are rate-limiting for the overall forward or reverse reaction. We can express the net reaction velocity v_{net} from substrate S to product P as the difference between the forward (from ES to EP) and the reverse reaction velocity (from EP to ES) (eq. 13.57):

$$v_{net} = k_f [ES] - k_r [EP] \qquad \text{eq. 13.57}$$

From here on, we can follow the standard procedure to obtain a velocity equation (Box 13.2). Division by the total enzyme concentration yields

$$\frac{v_{net}}{[E]_t} = \frac{k_f [ES] - k_r [EP]}{[E] + [ES] + [EP]} \qquad \text{eq. 13.58}$$

Expressing $[ES]$ and $[EP]$ in terms of the equilibrium dissociation constants K_S and K_P gives

$$\frac{v_{net}}{[E]_t} = \frac{k_f \frac{[S]}{K_S}[E] - k_r \frac{[P]}{K_P}[E]}{[E] + \frac{[S]}{K_S}[E] + \frac{[P]}{K_P}[E]} \qquad \text{eq. 13.59}$$

Elimination of $[E]$ on the right-hand side and multiplication with $[E]_t$ then gives v_{net} as

$$v_{net} = \frac{k_f [E]_t \frac{[S]}{K_S} - k_r [E]_t \frac{[P]}{K_P}}{1 + \frac{[S]}{K_S} + \frac{[P]}{K_P}} \qquad \text{eq. 13.60}$$

The terms $k_f[E]_t$ and $k_r[E]_t$ denote the maximum velocities for forward and reverse reactions, respectively, and eq. 13.60 can be written as

$$v_{net} = \frac{v_{max,f} \frac{[S]}{K_S} - v_{max,r} \frac{[P]}{K_P}}{1 + \frac{[S]}{K_S} + \frac{[P]}{K_P}} \qquad \text{eq. 13.61}$$

From eq. 13.61, we see that with increasing concentrations of product, the second term in the numerator increases, leading to a decrease of the numerator. In addition, the last term in the denominator, and thus the total denominator, increases. Both effects contribute to a decrease of the net reaction velocity v_{net} with increasing product concentration. This mechanism of enzyme inhibition is called *product inhibition*. Product inhibition depends on K_P, and is stronger for small values of K_P, *i.e.* when the enzyme binds tightly to the product. Strong product binding is often

observed in enzymes that catalyze reactions with chemically similar substrates and products. ATP-hydrolyzing enzymes frequently show product inhibition by ADP. Product inhibition is also very common in self-cleaving ribozymes where substrate and product only differ by one phosphodiester bond (see Chapter 16).

When the reaction reaches equilibrium, v_{net} is zero, and eq. 13.61 becomes

$$0 = v_{max,f}\frac{[S]_{eq}}{K_S} - v_{max,r}\frac{[P]_{eq}}{K_P} \qquad \text{eq. 13.62}$$

which can be rearranged to

$$K_{eq} = \frac{[P]_{eq}}{[S]_{eq}} = \frac{v_{max,f}K_P}{v_{max,r}K_S} \qquad \text{eq. 13.63}$$

K_{eq} is the equilibrium constant for the overall reaction from substrate to product. Note that K_{eq} only depends on the difference in free energy between product and substrate (see Section 2.1), and is not altered by the enzyme (Section 16.2). Using the expression in eq. 13.63 for K_{eq}, we can eliminate $v_{max,r}$ from eq. 13.61, and obtain an expression for v_{net} that is a function of $v_{max,f}$ only:

$$v_{net} = v_{max,f}\cdot\frac{[S]-\frac{[P]}{K_{eq}}}{K_S\left(1+\frac{[P]}{K_P}\right)+[S]} \qquad \text{eq. 13.64}$$

From eq. 13.64, we see that accumulated product leads to a higher apparent $K_{M,app} = K_S\cdot(1+[P]/K_P)$, which decreases v_{net}. In addition, v_{net} is reduced because the effective substrate concentration is reduced to $[S]-[P]/K_{eq}$ (Figure 13.14). The ratio $[P]/K_{eq}$ is equal to the concentration of the substrate in equilibrium, $[S]_{eq}$, and the difference between the substrate concentration $[S]$ and the ratio $[P]/K_{eq}$ in the numerator thus denotes the deviation of the substrate concentration from the equilibrium value. As long as $[S] > [S]_{eq}$, the net reaction velocity will remain positive, and the reaction will proceed from substrate to product. With ongoing forward reaction, the substrate concentration will approach the equilibrium value. For $[S]_{eq} > [S]$, on the other hand, v_{net} will be negative. In this case, the reaction proceeds from product to substrate, so as to establish the equilibrium concentration. This behavior is thus in accordance with the principle of Le Chatelier. At equilibrium, the numerator in eq. 13.64 becomes zero, and the net velocity is also zero. However, this does not mean that the reaction has stopped! The forward and reverse velocities can still be calculated from eq. 13.55 and eq. 13.56, but they are now equal: the reaction proceeds in both directions at the same speed. For this reason, we refer to chemical equilibria as "dynamic".

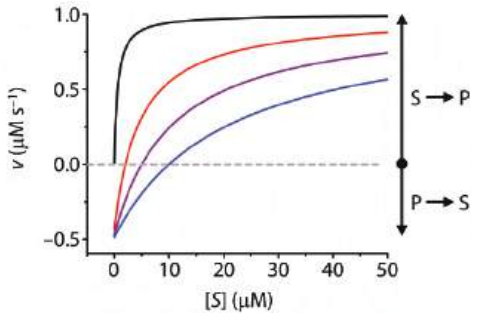
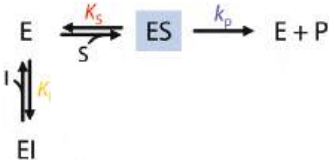


FIGURE 13.14: Product inhibition. Reaction velocity as a function of substrate concentration in the absence (black) and presence of 2 μM (red), 5 μM (purple), and 10 μM (blue) product; K_S = 0.6 μM, K_P = 0.3 μM, $v_{max,f}$ = 1 μM s^{-1}, K_{eq} = 1. For positive net velocities, the reaction proceeds toward products; for negative velocities, it proceeds toward substrates.

If substrate and product of an enzyme are chemically rather similar, the product efficiently competes with the substrate for the substrate binding site. This type of inhibition is therefore also termed competitive inhibition. We will treat this and other inhibition mechanisms and the effects on k_{cat} and K_M in the following (Sections 13.6.2–13.6.4).

13.6.2 Competitive inhibition

A *competitive inhibitor* competes with the substrate for the same binding site on the enzyme. Thus, we have a substrate binding equilibrium (K_S), and a competing equilibrium of inhibitor binding (K_I). Only the ES complex, but not the enzyme-inhibitor complex EI, can convert the substrate into products (Scheme 13.10).



scheme 13.10

As only the enzyme-substrate complex ES contributes to product formation, we can express the reaction velocity as

$$v = k_P[ES] \qquad \text{eq. 13.3}$$

We now follow the standard procedure (Box 13.2) and divide by the total enzyme concentration, and take into account that the enzyme can exist as free enzyme E, as ES complex, or as EI complex.

$$\frac{v}{[E]_t} = \frac{k_P[ES]}{[E]+[ES]+[EI]} \qquad \text{eq. 13.65}$$

With [*ES*] expressed in terms of the equilibrium dissociation constant of ES, K_S, and [*EI*] expressed in terms of the equilibrium constant for dissociation of the EI complex, K_I, we obtain

$$\frac{v}{[E]_t} = \frac{k_P\frac{[S]}{K_S}[E]}{[E]+\frac{[S]}{K_S}[E]+\frac{[I]}{K_I}[E]} \qquad \text{eq. 13.66}$$

With $v_{max} = k_p[E]_t$, this equation can be simplified to

$$\frac{v}{v_{max}} = \frac{[S]}{K_S\left(1+\frac{[I]}{K_I}\right)+[S]} = \frac{[S]}{K_{M,app}+[S]} \qquad \text{eq. 13.67}$$

with

$$K_{M,app} = K_S\left(1+\frac{[I]}{K_I}\right) \qquad \text{eq. 13.68}$$

A competitive inhibitor thus does not affect the v_{max} of the reaction, but the apparent Michaelis constant is increased by $(1+[I]/K_I)$ compared to K_S (Figure 13.15). At increasing substrate concentrations, the substrate can compete with the inhibitor for binding, and v_{max} is reached.

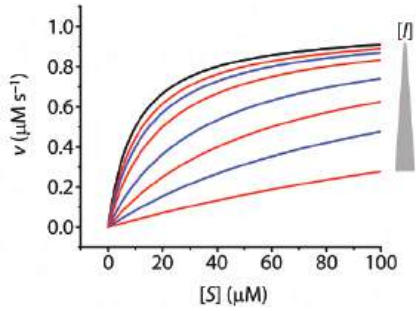


FIGURE 13.15: Competitive inhibition. Reaction velocity as a function of the substrate concentration for K_S = 10 μM, v_{max} = 1 μM s^{-1}, and K_I = 50 μM in the absence of inhibitor (black), and in the presence of 1, 10, 20, 50, 100, 200, and 500 μM inhibitor (red: 1, 20, 100, and 500 μM, blue: 10, 50, and 200 μM). A competitive inhibitor interferes with substrate binding, and $K_{M,app}$ is higher than K_S. At saturation with substrate, v_{max} is reached.

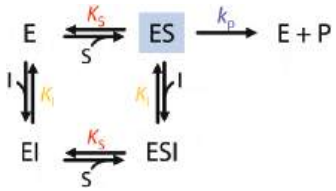
Many enzymes are inhibited competitively by substrate analogs. One example is succinate dehydrogenase, the enzyme that catalyzes the oxidation of succinate to fumarate in the citric acid cycle, which is inhibited by malonate. Malonate binds to the substrate binding site, but is not turned over.

Competitive inhibition can also be exploited therapeutically. The enzyme dihydrofolate reductase (DHFR) catalyzes the reduction of dihydrofolate to tetrahydrofolate, which is the central cofactor for methyl group transfer in the synthesis of purine nucleotides and of thymidylate. The DHFR reaction is particularly important for proliferating cells, such as cancer cells. In chemotherapy, the competitive DHFR inhibitor methotrexate is used to inhibit DHFR in cancer cells. The bacterial DHFR is also a drug target. Bacteria have a different type of DHFR that is competitively inhibited by trimethoprim. This antibiotic has little effect on the human enzyme.

Non-hydrolyzable ATP analogs can act as competitive inhibitors for enzymes that hydrolyze ATP. Novobiocin, an inhibitor of DNA gyrase that is used in the treatment of bacterial infections, inhibits the DNA supercoiling activity (see Section 19.5.3.2) of gyrase by competing with ATP for binding to the ATP binding site.

13.6.3 Non-competitive inhibition

In another mode of enzyme inhibition, the inhibitor I binds to both the free enzyme and to the enzyme-substrate complex, leading to formation of an EI or ESI complex (Scheme 13.11). Here, the inhibitor binds to a different site on the enzyme from that of the substrate. We can assume that K_S is the same for the ES and ESI complexes and that ESI does not catalyze product formation:



scheme 13.11

The reaction velocity is

$$v = k_p[ES] \qquad \text{eq. 13.3}$$

leading to a similar expression as for the case of competitive inhibition (eq. 13.65):

$$\frac{v}{[E]_t} = \frac{k_p[ES]}{[E]+[ES]+[EI]+[ESI]} \qquad \text{eq. 13.69}$$

The only difference from eq. 13.65 is the appearance of [*ESI*] in the denominator. With [*ES*], [*EI*], and [*ESI*] expressed in terms of the equilibrium dissociation constants K_S and K_I, we obtain

$$\frac{v}{[E]_t} = \frac{k_p \frac{[E][S]}{K_S}}{[E] + \frac{[E][S]}{K_S} + \frac{[E][I]}{K_I} + \frac{[E][S][I]}{K_S K_I}} \qquad \text{eq. 13.70}$$

Again, we can cancel [*E*] from all terms, divide eq. 13.70 by k_p, and substitute $k_p[E]_t$ by v_{max}. Expansion of the right-hand side with K_S and simplifying the denominator leads to

$$\frac{v}{v_{max}} = \frac{[S]}{K_S\left(1+\frac{[I]}{K_I}\right)+[S]\left(1+\frac{[I]}{K_I}\right)} \qquad \text{eq. 13.71}$$

After transferring the factor $(1+[I]/K_I)$ to the left-hand side, we can express eq. 13.71 in the form

$$\frac{v}{v_{max,app}} = \frac{[S]}{K_S+[S]} \qquad \text{eq. 13.72}$$

with

$$v_{max,app} = \frac{v_{max}}{\left(1+\frac{[I]}{K_I}\right)} \qquad \text{eq. 13.73}$$

Thus, by sequestering the enzyme in an inactive EI or ESI complex, a *non-competitive inhibitor* reduces v_{max} by $(1+[I]/K_I)$ but does not affect K_M (Figure 13.16). In other words, a non-competitive inhibitor does not affect substrate binding, but only reduces the reaction rate.

Non-competitive inhibitors bind to sites on the enzyme distinct from the substrate binding site, and interfere indirectly with substrate binding. This type of inhibition is often found in feedback inhibition, and is referred to as *allosteric* (Greek: another place). An example is the inhibition of pyruvate kinase by alanine. Pyruvate kinase catalyzes the oxidative dephosphorylation of phosphoenol pyruvate to pyruvate, which can be converted to alanine (see Figure 8.3). Alanine acts as a non-competitive inhibitor of pyruvate kinase, ensuring that pyruvate production is reduced when sufficient amounts of alanine are present. Allosteric activation is also common.

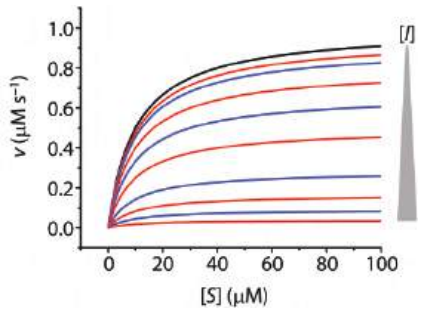
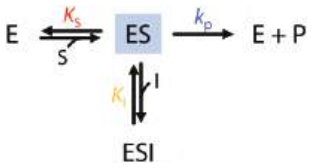


FIGURE 13.16: Non-competitive inhibition. Reaction velocity as a function of the substrate concentration for K_S = 10 μM, v_{max} = 1 μM s^{-1}, and K_I = 50 μM in the absence of inhibitor (black), and in the presence of increasing concentrations of a non-competitive inhibitor (red: 1, 5, 20, 100, and 500 μM, blue: 2, 10, 50, and 200 μM). A non-competitive inhibitor sequesters the enzyme in an inactive complex, leading to a reduced v_{max}. The inhibitor does not affect substrate binding, however, and $K_{M,app}$ equals K_S.

13.6.4 Mixed inhibition

We have derived eq. 13.72 and eq. 13.73 for non-competitive inhibition assuming that K_S for dissociation of substrate from ES and ESI is identical. However, these constants often differ, leading to altered K_M and v_{max}. This type of inhibition is called *mixed inhibition*.

One type of mixed inhibition is *uncompetitive inhibition*. In this case, the inhibitor binds to ES, but not to the free enzyme, *i.e.* the binding site for the inhibitor is created only upon binding of substrate to the enzyme. Scheme 13.11 then simplifies to Scheme 13.12:



scheme 13.12

We can now express the reaction velocity as the product of k_p and the concentration of active enzyme, *i.e.* the ES complex, divide the equation by the total enzyme concentration, express all complex concentrations by [*E*], [*S*], and K_S or [*E*], [*I*], and K_I, substitute $k_p[E]_t$ by v_{max}, and simplify to

$$\frac{v}{v_{max,app}} = \frac{[S]}{K_{M,app} + [S]} \qquad \text{eq. 13.74}$$

with

$$v_{max,app} = \frac{v_{max}}{\left(1 + \frac{[I]}{K_I}\right)} \qquad \text{eq. 13.73}$$

and

$$K_{M,app} = \frac{K_S}{1 + \frac{[I]}{K_I}} \qquad \text{eq. 13.75}$$

An uncompetitive inhibitor thus reduces v_{max} and K_M by a factor of $(1+[I]/K_I)$ (Figure 13.17). The decreased $K_{M,app}$ in non-competitive inhibition corresponds to favored substrate binding when the inhibitor is present. Note that this factor $(1+[I]/K_I)$ is equal to the factor by which $K_{M,app}$ is increased over K_S by a competitive inhibitor that competes with the binding, whereas the uncompetitive inhibitor favors substrate binding by binding to the product of the substrate binding equilibrium ES (principle of Le Chatelier; Section 2.1.2).

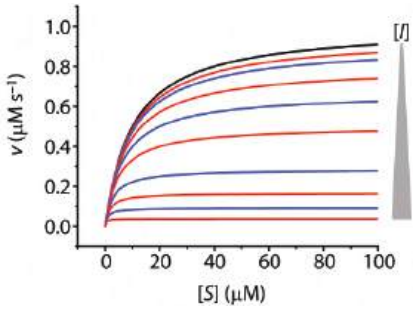


FIGURE 13.17: Uncompetitive inhibition. Reaction velocity as a function of the substrate concentration for K_S = 10 μM, v_{max} = 1 μM s^{-1}, and K_I = 50 μM in the absence of inhibitor (black), and in the presence of 1, 2, 5, 10, 20, 50, 100, 200, and 500 μM of a uncompetitive inhibitor (red: 1, 5, 20, 100, and 500 μM, blue: 2, 10, 50, and 200 μM). An uncompetitive inhibitor sequesters the enzyme in an inactive complex, leading to a reduced v_{max}. The inhibitor also affects substrate binding by formation of ESI, leading to a decreased $K_{M,app}$ compared to K_S.

We have seen that different types of inhibition differentially affect v_{max} and K_M. The inhibition types cause characteristic differences in reciprocal plots, and from these plots rapid classification of inhibitors is possible (Box 13.4).

BOX 13.4: ENZYME INHIBITION AND LINEAR PLOTS.

Reciprocal plots after Lineweaver-Burk or Eadie-Hofstee (Section 13.2; eq. 13.21, eq. 13.22) are particularly useful to study inhibition (Figure 13.18). Competitive inhibition reduces K_M, but does not affect v_{max}. Lineweaver-Burk plots for enzymes in the presence of increasing concentrations of inhibitor have the same y-axis intercept, but the slope increases with the inhibitor concentration. The same behavior is seen in an Eadie-Hofstee plot. Non-competitive inhibition reduces v_{max}, but not K_M. This type of inhibition is reflected in an increased slope and y-axis intercept with increasing inhibitor concentration in Lineweaver-Burk plots, and a decreasing y-axis intercept, but identical slopes, in Eadie-Hofstee plots. For an uncompetitive inhibitor, both v_{max} and K_M are affected, and slopes and intercepts change with increasing inhibitor concentrations in both plots.

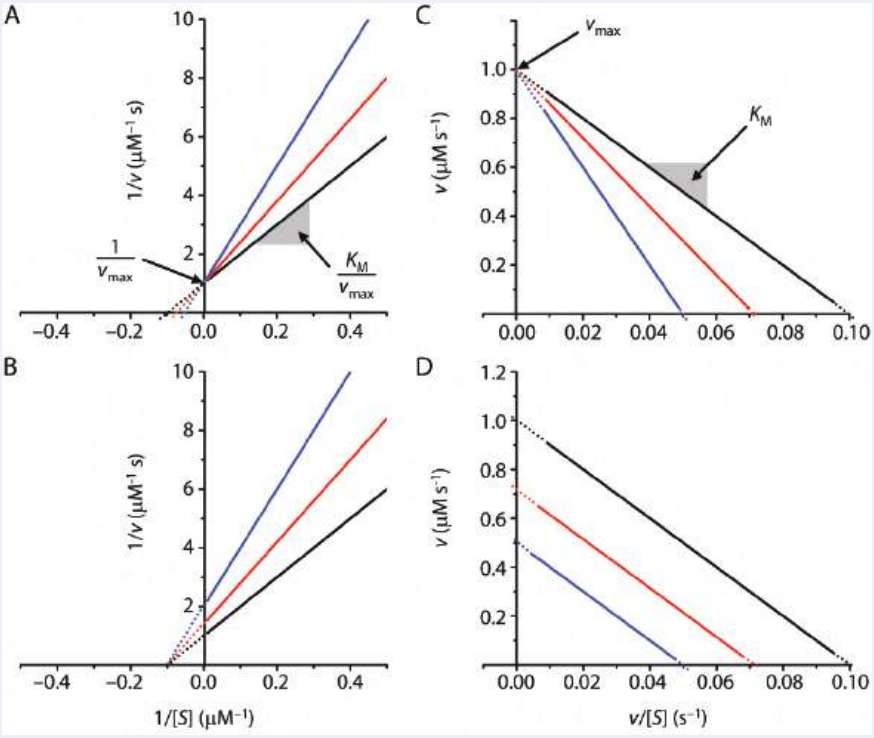


FIGURE 13.18: Reciprocal plots and inhibition. Lineweaver-Burk (A,B) and Eadie-Hofstee plots (C,D) for competitive inhibition (A,C) and non-competitive inhibition (B,D) in the absence of inhibitor (black), and in the presence of increasing concentrations of inhibitor (red, blue).

13.6.5 Quantifying enzyme inhibition: inhibition constants *versus* IC_{50} values

In the previous sections, we have quantified the effect of inhibitors that inhibit enzyme activity by binding reversibly to the enzyme and/or the ES complex through the value of K_I, the dissociation constant of the EI or ESI complex (Schemes 13.1, 13.11, and 13.12). The identification of enzyme inhibitors that inhibit the reaction catalyzed with high efficiency is frequently characterized by the IC_{50} value, the concentration of the inhibitor at which the enzymatic reaction is reduced to 50% of the value in the absence of the inhibitor. The reaction velocity v in the absence of the inhibitor follows the Michaelis-Menten equation (eq. 13.8):

$$v = v_{max} \cdot \frac{[S]}{K_S + [S]} \qquad \text{eq. 13.8}$$

In the presence of a competitive inhibitor at the concentration $[I]$, the velocity v_I is reduced to

$$v_I = v_{max} \cdot \frac{[S]}{K_S\left(1 + \frac{[I]}{K_I}\right) + [S]} \qquad \text{eq. 13.67, eq. 13.68}$$

At $[I] = IC_{50}$, v_I equals 0.5 v:

$$v_I = 0.5v \qquad \text{eq. 13.76}$$

which gives

$$\frac{[S]}{K_S\left(1 + \frac{IC_{50}}{K_I}\right) + [S]} = 0.5\frac{[S]}{K_S + [S]} \qquad \text{eq. 13.77}$$

Inversion of the fractions on both sides leads to

$$\frac{K_S\left(1 + \frac{IC_{50}}{K_I}\right) + [S]}{[S]} = \frac{K_S + [S]}{0.5[S]} \qquad \text{eq. 13.78}$$

which equals

$$K_S\left(1 + \frac{IC_{50}}{K_I}\right) + [S] = 2K_S + 2[S] \qquad \text{eq. 13.79}$$

or

$$K_S + K_S\frac{IC_{50}}{K_I} + [S] = 2K_S + 2[S] \qquad \text{eq. 13.80}$$

Simplification of eq. 13.80 results in the relationship

$$K_S\frac{IC_{50}}{K_I} = K_S + [S] \qquad \text{eq. 13.81}$$

which can be solved for IC_{50}:

$$IC_{50} = K_I \frac{K_S + [S]}{K_S} = K_I \left(1 + \frac{[S]}{K_S}\right) \qquad \text{eq. 13.82}$$

The expression we derived for the relationship between IC_{50} and K_I contains two important messages: first, eq. 13.82 tells us that IC_{50} approaches K_I only for substrate concentrations much smaller than K_S, when the term in parentheses approaches unity (Figure 13.19). In practice, IC_{50} values for competitive inhibitors will therefore not reach K_I because reaction velocities at substrate concentrations far below K_S are very small, and cannot be measured reliably. Second, IC_{50} and K_I are directly proportional for constant values of the substrate concentration $[S]$ and the dissociation constant of the enzyme-substrate complex K_S. Provided that the substrate concentration is kept constant and the same substrate with the same K_S is used in different experiments assessing different competitive inhibitors, IC_{50} values can be used as a proxy for inhibition constants (Figure 13.19). Only under these conditions is it allowed to compare IC_{50} values for different inhibitors. As soon as the substrate concentration is altered, however, the IC_{50} values are not comparable anymore (Figure 13.19). The same applies if inhibition of the enzymatic reaction is tested using different substrates with different K_S values: as soon as K_S changes, the IC_{50} value is altered as well. While these limitations are not necessarily a problem, it is important to be aware of them when interpreting IC_{50} values: IC_{50} values are not equilibrium constants, but apparent values that can only be interpreted in comparison with other values determined under identical conditions! The situation is different for non-competitive inhibitors, where $IC_{50} = K_I$.

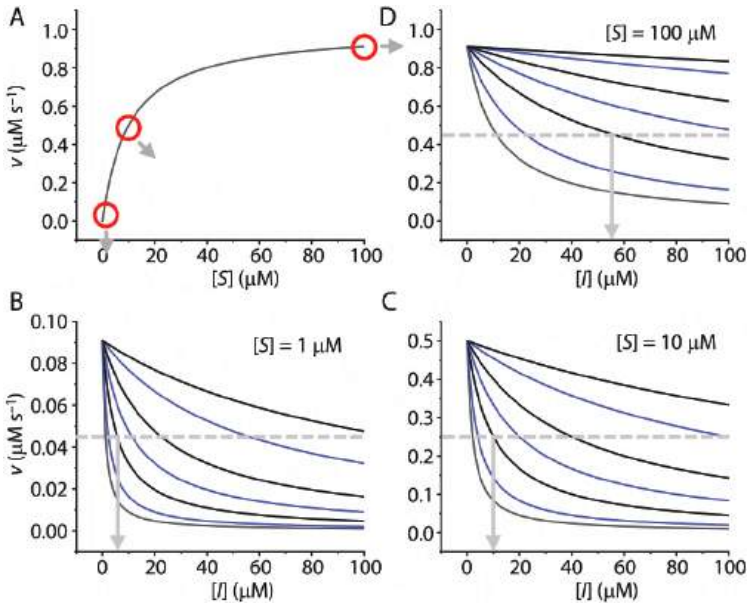


FIGURE 13.19: Inhibition constants and IC_{50} values. A: Reaction velocity as a function of substrate concentration for an enzyme with K_S = 10 µM and v_{max} = 1 µM s^{-1}. The red circles highlight the substrate concentrations used for calculation of the inhibition curves at different concentrations of a competitive inhibitor in B-D. B: Reaction velocity v as a function of inhibitor concentration at $[S]$ = 1 µM for inhibitors with different K_I values (1, 2, 5, 10, 20, 50, 100 µM, from bottom to top). C: Reaction velocity v as a function of inhibitor concentration at $[S]$ = 10 µM for inhibitors with different K_I values. D: Reaction velocities as a function of inhibitor concentration at $[S]$ = 100 µM. The broken gray lines in B-D mark the half-maximal reaction velocities. K_I values are 1 µM, 2 µM, 5 µM, 10 µM, 20 µM, 50 µM, or 100 µM, from bottom to top. The gray arrows indicate the IC_{50} values determined for an inhibitor with K_I = 5 µM. The IC_{50} values for this inhibitor are IC_{50} = 5.5 µM at $[S]$ = 1 µM, IC_{50} = 10 µM at $[S]$ = 10 µM, and IC_{50} = 55 µM at $[S]$ = 100 µM.

13.6.6 Time-dependent inhibition

The derivations of reaction velocities in the presence of competitive and non-competitive inhibitors were all made with the assumption that the binding equilibria between enzyme, inhibitor, and enzyme-inhibitor complex(es) are rapid compared to product formation by the enzyme. As a result, the formation of product occurs at constant velocity, and the concentration of product increases linearly with the reaction time. However, if the interaction of the inhibitor with the enzyme is slow, then this may not be the case (Scheme 13.13). In one possible scenario, the inhibitor binds in two steps (see Chapter 11). The first step is the rapid formation of a collision complex EI; this collision complex then rearranges reversibly, but slowly, to the inactive form EI*.



scheme 13.13

On incubation of an enzyme with its substrate in the presence of such slowly-interacting inhibitors, the formation of product does not depend linearly on the time. Instead, the reaction velocity decreases with time, hence the name *time-dependent* or *slow-onset inhibitors* (Figure 13.20). Time-dependent inhibition is frequently observed with high-affinity inhibitors, as a high affinity for the enzyme is associated with a small rate constant of dissociation (k_{-2} in Scheme 13.13). Many medicinally relevant compounds that bind tightly, often covalently, to their targets show time-dependent inhibition.

We will see in the following that we can derive the velocity equation for the enzymatic reaction in the presence of such time-dependent inhibitors by the standard procedure (see Box 13.2), with the appropriate modifications to take into account the time-dependence. We start with the general expression for the reaction velocity v as :

$$v = k_p[ES] \qquad \text{eq. 13.83}$$

As always, we then divide both sides of the equation by the total enzyme concentration:

$$\frac{v}{[E]_t} = \frac{k_p[ES]}{[E]+[ES]+[EI]+[EI^*]} \qquad \text{eq. 13.84}$$

In the standard procedure, we would now express all complexes in the denominator in terms of $[E]$ using the definition of the equilibrium constants. The equilibrium constants K_S and K_I are defined as

$$K_S = \frac{[E][S]}{[ES]} \qquad \text{eq. 13.85}$$

and

$$K_I = \frac{[E][I]}{[EI]} \qquad \text{eq. 13.86}$$

We can express the equilibrium constant K_{slow} for the slow equilibrium between EI and EI* using the forward and reverse rate constants k_2 and k_{-2}:

$$K_{slow} = \frac{[EI^*]}{[EI]} = \frac{k_2}{k_{-2}} \qquad \text{eq. 13.87}$$

However, due to the slow equilibration of EI and EI*, both concentrations will be time-dependent during the time course of the actual experiment. Hence, we need to deviate from the standard

procedure here, and express these changes by differential equations. The change in the concentration of EI with time is

$$\frac{\mathrm{d}[EI]}{\mathrm{d}t} = -k_2[EI] + k_{-2}[EI^*] \qquad \text{eq. 13.88}$$

By solving eq. 13.85 for $[E]$ and substituting the expression obtained in the definition of K_I (eq. 13.86), we obtain for the relationship between EI and ES:

$$[EI] = \frac{K_S}{[S]}\frac{[I]}{K_I}[ES] \qquad \text{eq. 13.89}$$

or

$$\frac{\mathrm{d}[EI]}{\mathrm{d}t} = \frac{K_S}{[S]}\frac{[I]}{K_I}\frac{\mathrm{d}[ES]}{\mathrm{d}t} \qquad \text{eq. 13.90}$$

Now we combine eq. 13.90 with the other differential equation (eq. 13.88):

$$\frac{\mathrm{d}[EI]}{\mathrm{d}t} = -k_2[EI] + k_{-2}[EI^*] = \frac{K_S}{[S]}\frac{[I]}{K_I}\frac{\mathrm{d}[ES]}{\mathrm{d}t} \qquad \text{eq. 13.91}$$

and solve eq. 13.91 for $[EI^*]$, which gives

$$[EI^*] = \frac{1}{k_{-2}}\left(\frac{K_S}{[S]}\frac{[I]}{K_I}\frac{d[ES]}{dt} + k_2[EI]\right) \qquad \text{eq. 13.92}$$

By substituting eq. 13.89 for $[EI]$, the expression becomes

$$[EI^*] = \frac{1}{k_{-2}}\left(\frac{K_S}{[S]}\cdot\frac{[I]}{K_I}\cdot\frac{\mathrm{d}[ES]}{\mathrm{d}t} + k_2\frac{K_S}{[S]}\cdot\frac{[I]}{K_I}[ES]\right) \qquad \text{eq. 13.93}$$

or

$$[EI^*] = \frac{K_S}{[S]}\cdot\frac{[I]}{K_I}\cdot\frac{1}{k_{-2}}\left(\frac{\mathrm{d}[ES]}{\mathrm{d}t} + k_2[ES]\right) \qquad \text{eq. 13.94}$$

Next, we return to the mass conservation for the enzyme

$$[E]_t = [E] + [ES] + [EI] + [EI^*] \qquad \text{eq. 13.95}$$

Instead of expressing all enzyme species in terms of $[E]$, as we do in the standard procedure, we now express these concentrations in terms of $[ES]$ by using the definitions of K_S (eq. 13.85) and K_I (eq. 13.86) as well as eq. 13.94:

$$[E]_t = \frac{K_S}{[S]}[ES] + [ES] + \frac{K_S}{[S]}\cdot\frac{[I]}{K_I}[ES] + \frac{K_S}{[S]}\cdot\frac{[I]}{K_I}\cdot\frac{1}{k_{-2}}\left(\frac{\mathrm{d}[ES]}{\mathrm{d}t} + k_2[ES]\right) \qquad \text{eq. 13.96}$$

Solving for $\mathrm{d}[ES]$ and separating the variables gives

$$[E]_t - \frac{K_S}{[S]}[ES] = [ES] + \frac{K_S}{[S]}\cdot\frac{[I]}{K_I}[ES] + \frac{K_S}{[S]}\cdot\frac{[I]}{K_I}\cdot\frac{1}{k_{-2}}\left(\frac{\mathrm{d}[ES]}{\mathrm{d}t} + k_2[ES]\right) \qquad \text{eq. 13.97}$$

and

$$[E]_t - \frac{K_S}{[S]}[ES] - [ES] - \frac{K_S}{[S]} \cdot \frac{[I]}{K_I}[ES] = \frac{K_S}{[S]} \cdot \frac{[I]}{K_I} \cdot \frac{1}{k_{-2}} \left(\frac{d[ES]}{dt} + k_2[ES] \right) \quad \text{eq. 13.98}$$

or

$$\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} \cdot \left([E]_t - \frac{K_S}{[S]}[ES] - [ES] - \frac{K_S}{[S]} \cdot \frac{[I]}{K_I}[ES] \right) = \frac{d[ES]}{dt} + k_2[ES] \quad \text{eq. 13.99}$$

or

$$\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} \cdot \left([E]_t - \frac{K_S}{[S]}[ES] - [ES] - \frac{K_S}{[S]} \cdot \frac{[I]}{K_I}[ES] \right) - k_2[ES] = \frac{d[ES]}{dt} \quad \text{eq. 13.100}$$

By multiplying the terms in the parentheses with the pre-factor and canceling a number of factors, we obtain

$$\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2}[E]_t - \frac{K_I}{[I]} \cdot k_{-2}[ES] - \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2}[ES] - k_{-2}[ES] - k_2[ES] = \frac{d[ES]}{dt} \quad \text{eq. 13.101}$$

which can be rearranged to

$$\frac{d[ES]}{dt} = \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2}[E]_t - \left(\frac{K_I}{[I]} \cdot k_{-2} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} + k_{-2} + k_2 \right)[ES] \quad \text{eq. 13.102}$$

and further to

$$d[ES] \frac{1}{\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2}[E]_t - \left(\frac{K_I}{[I]} \cdot k_{-2} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} + k_{-2} + k_2 \right)[ES]} = dt \quad \text{eq. 13.103}$$

Before we integrate eq. 13.103, we first define the variables a and b as

$$a = \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2}[E]_t \quad \text{eq. 13.104}$$

and

$$b = \frac{K_I}{[I]} \cdot k_{-2} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} + k_{-2} + k_2 \quad \text{eq. 13.105}$$

which simplifies eq. 13.103 to

$$d[ES] \frac{1}{a - b[ES]} = dt \quad \text{eq. 13.106}$$

Eq. 13.106 can be integrated by substitution (see Section 31.6), with

$$u = a - b[ES] \quad \text{eq. 13.107}$$

and

$$\frac{du}{d[ES]} = -b \quad \text{eq. 13.108}$$

or

$$\mathrm{d}[ES] = -\frac{\mathrm{d}u}{b} \qquad \text{eq. 13.109}$$

Substituting into eq. 13.106 gives

$$-b\frac{\mathrm{d}u}{u} = \mathrm{d}t \qquad \text{eq. 13.110}$$

which can directly be integrated to

$$-\frac{1}{b}\ln u = t + \text{const.} \qquad \text{eq. 13.111}$$

Re-substitution gives

$$-\frac{1}{b}\ln\left(a - b[ES]\right) = t + \text{const.} \qquad \text{eq. 13.112}$$

We can determine the integration constant *const.* with the boundary condition that the concentration of ES at $t = 0$ is $[ES]_0$ – we will determine a value for this concentration later. With

$$-\frac{1}{b}\ln\left(a - b[ES]_0\right) = \text{const.} \qquad \text{eq. 13.113}$$

eq. 13.112 becomes

$$-\frac{1}{b}\ln\left(a - b[ES]\right) = t - \frac{1}{b}\ln\left(a - b[ES]_0\right) \qquad \text{eq. 13.114}$$

which we can rearrange to

$$-\frac{1}{b}\ln\left(\frac{a - b[ES]}{a - b[ES]_0}\right) = t \qquad \text{eq. 13.115}$$

or

$$\ln\left(\frac{a - b[ES]}{a - b[ES]_0}\right) = -bt \qquad \text{eq. 13.116}$$

Applying the exponential function on both sides then gives

$$\frac{a - b[ES]}{a - b[ES]_0} = e^{-bt} \qquad \text{eq. 13.117}$$

or

$$a - b[ES] = \left(a - b[ES]_0\right) \cdot e^{-bt} \qquad \text{eq. 13.118}$$

Solving for $[ES]$ leads to

$$[ES] = \frac{a}{b} - \left(\frac{a}{b} - [ES]_0\right) \cdot e^{-bt} \qquad \text{eq. 13.119}$$

For $t \to 0$, the exponential function in eq. 13.119 approaches a value of unity, and $[ES]$ equals $[ES]_0$, commensurate with our boundary condition. For $t \to \infty$, e^{-bt} becomes zero, and eq. 13.119 simplifies to

$$[ES]_\infty = \frac{a}{b} \qquad \text{eq. 13.120}$$

With re-substitution of a and b from eq. 13.104 and eq. 13.105 we obtain

$$[ES]_\infty = \frac{\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} [E]_t}{\frac{K_I}{[I]} \cdot k_{-2} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot k_{-2} + k_{-2} + k_2} \qquad \text{eq. 13.121}$$

By reducing by k_{-2}

$$[ES]_\infty = \frac{\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot [E]_t}{\frac{K_I}{[I]} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} + 1 + \frac{k_2}{k_{-2}}} = \frac{\frac{[S]}{K_S} \cdot \frac{K_I}{[I]} \cdot [E]_t}{\frac{K_I}{[I]} + \frac{[S]}{K_S} \cdot \frac{K_I}{[I]} + \frac{k_{-2} + k_2}{k_{-2}}} \qquad \text{eq. 13.122}$$

expanding with $[I]$

$$[ES]_\infty = \frac{\frac{[S]}{K_S} \cdot K_I \cdot [E]_t}{K_I + \frac{[S]}{K_S} \cdot K_I + [I] \frac{k_{-2} + k_2}{k_{-2}}} \qquad \text{eq. 13.123}$$

reducing by K_I

$$[ES]_\infty = \frac{\frac{[S]}{K_S} \cdot [E]_t}{1 + \frac{[S]}{K_S} + \frac{[I]}{K_I} \cdot \frac{k_{-2} + k_2}{k_{-2}}} \qquad \text{eq. 13.124}$$

and expanding the fraction with K_S, we can rearrange eq. 13.121 to a Michaelis-Menten-like equation:

$$[ES]_\infty = [E]_t \cdot \frac{[S]}{K_S + [S] + K_S \cdot \frac{[I]}{K_I} \cdot \frac{k_{-2} + k_2}{k_{-2}}} = [E]_t \cdot \frac{[S]}{K_S \left(1 + \frac{[I]}{K_I} \cdot \frac{k_{-2} + k_2}{k_{-2}}\right) + [S]} \qquad \text{eq. 13.125}$$

With

$$K_I \cdot \frac{k_{-2}}{k_{-2} + k_2} = K_I^* \qquad \text{eq. 13.126}$$

eq. 13.125 becomes

$$[ES]_\infty = [E]_t \cdot \frac{[S]}{K_S \left(1 + \frac{[I]}{K_I^*}\right) + [S]} \qquad \text{eq. 13.127}$$

where K_I^* is the apparent inhibition constant describing the apparent equilibrium between E and I and EI*:

$$K_I^* = \frac{[E][I]}{[EI]+[EI^*]} \qquad \text{eq. 13.128}$$

Note that, according to eq. 13.126, K_I^* is significantly different from K_I only if $k_2 > k_{-2}$. This condition is in agreement with the requirement for a small k_{-2} for time-dependent inhibition to occur. If $k_{-2} > k_2$, then the equilibrium between EI and EI* is on the side of EI, and classical competitive inhibition is observed.

Returning to the time dependence of [*ES*] (eq. 13.119), we see that the parameter *b* in the exponent of this expression corresponds to a rate constant that describes the decrease in the concentration of the ES complex, *i.e.* the inactivation of the enzyme over time through the slow equilibration of EI and EI*. Knowing [*ES*](*t*), we can now finally derive the expression for the reaction velocity $v(t) = k_p[ES](t)$ as

$$v(t) = k_p\left(\frac{a}{b} - \left(\frac{a}{b} - [ES]_0\right)\cdot e^{-bt}\right) \qquad \text{eq. 13.129}$$

which can be rearranged to

$$v(t) = k_p\frac{a}{b} - k_p\frac{a}{b}\cdot e^{-bt} + k_p[ES]_0\cdot e^{-bt} \qquad \text{eq. 13.130}$$

or

$$v(t) = k_p\frac{a}{b}\left(1 - e^{-bt}\right) + k_p[ES]_0\cdot e^{-bt} \qquad \text{eq. 13.131}$$

By substituting $[ES]_\infty$ for a/b (eq. 13.120), we obtain

$$v(t) = k_p[ES]_\infty\left(1 - e^{-bt}\right) + k_p[ES]_0\cdot e^{-bt} \qquad \text{eq. 13.132}$$

With $k_p[ES]_0 = v_0$ and $k_p[ES]_\infty = v_s$, eq. 13.132 becomes

$$v(t) = v_s\left(1 - e^{-bt}\right) + v_0\cdot e^{-bt} \qquad \text{eq. 13.133}$$

or

$$v(t) = v_s + \left(v_0 - v_s\right)\cdot e^{-bt} \qquad \text{eq. 13.134}$$

This velocity equation describes the reaction velocity *v* from the beginning of the experiment, when the enzyme is mixed with substrate and inhibitor, and the equilibrium between E and S as well as E and I is established instantaneously (v_0), until also EI and EI* have equilibrated, and the reaction velocity has reached its steady-state value (v_s). The drop in reaction velocity from v_0 at $t = 0$ to the final value v_s follows a single-exponential function, with the first-order rate constant *b* that describes the inactivation of the enzyme by the slow equilibration of EI to EI*. The value for *b* in turn depends on the rate constants for the slow inactivation step, governed by k_2 and k_{-2}, on the equilibrium constants K_I and K_S, and on the substrate and inhibitor concentrations, [*S*] and [*I*] (eq. 13.105, Figure 13.20). The inverse of the rate constant *b*, *i.e.* the lifetime, shows a hyperbolic dependence on the inhibitor concentration (Figure 13.20).

We can calculate v_0 and v_s in eq. 13.134 following the relationship between the reaction velocity v and the concentration of the product-forming ES complex as

$$v_0 = k_p[ES]_0 = k_p[E]_t \cdot \frac{[S]}{K_S\left(1+\frac{[I]}{K_I}\right)+[S]} \qquad \text{eq. 13.135}$$

and

$$v_s = k_p[ES]_\infty = k_p[E]_t \cdot \frac{[S]}{K_S\left(1+\frac{[I]}{K_I^*}\right)+[S]} \qquad \text{eq. 13.136}$$

Note that eq. 13.135 is the equation for the reaction velocity in the presence of a competitive inhibitor (see eq. 13.67): at $t = 0$, when EI* has not been formed yet due to the slow equilibration between EI and EI*, Scheme 13.13 simplifies to the scheme for competitive inhibition (see Scheme 13.10).

For $t \to \infty$, the interconversion of EI and EI* is at equilibrium. The reaction velocity then reaches its minimum value v_s. For the extreme case of irreversibly binding inhibitors ($k_{-2} = 0$, $K_I^* = 0$; eq. 13.126), v_s becomes zero.

From eq. 13.134, we can now directly calculate the integrated rate law, *i.e.* the product concentration as a function of time, by integration:

$$[P](t) = \int v(t)\,dt = \int v_S\,dt + \int (v_0 - v_S)\cdot e^{-bt}\,dt \qquad \text{eq. 13.137}$$

Separate evaluation of the two integrals gives

$$[P](t) = v_S t - \frac{v_0 - v_S}{b}\cdot e^{-bt} + \text{const.} \qquad \text{eq. 13.138}$$

With the boundary condition $[P](0) = 0$, *i.e.* no product has been formed at $t = 0$, we can determine the integration constant as

$$\text{const.} = \frac{v_0 - v_S}{b} \qquad \text{eq. 13.139}$$

which gives

$$[P](t) = v_S t - \frac{v_0 - v_S}{b}\cdot e^{-bt} + \frac{v_0 - v_S}{b} \qquad \text{eq. 13.140}$$

or

$$[P](t) = \frac{v_0 - v_S}{b}\cdot\left(1 - e^{-bt}\right) + v_S t \qquad \text{eq. 13.141}$$

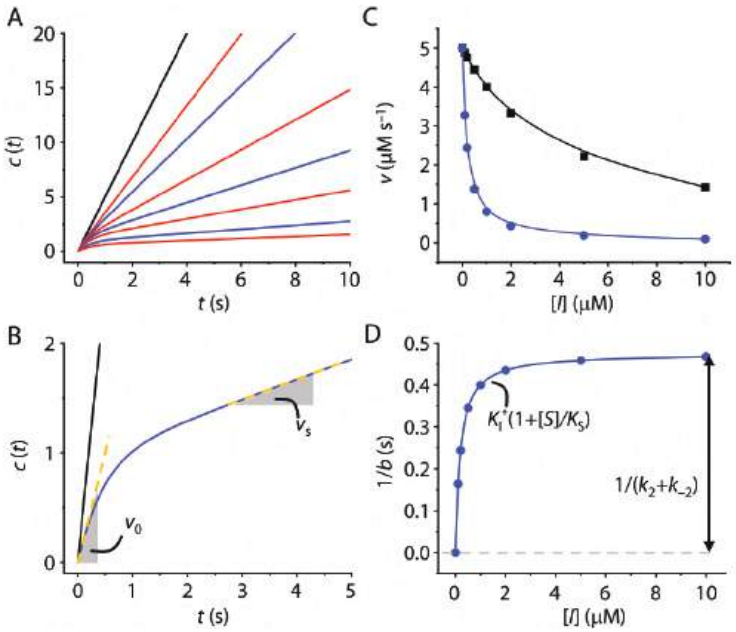


FIGURE 13.20: Time-dependent enzyme inhibition. A: Time-dependence of the appearance of product P in the presence of increasing concentrations of a time-dependent inhibitor I, following Scheme 13.13. K_S = 5 μM, K_I = 2 μM, v_{max} = 10 μM s^{-1}, k_2 = 2 s^{-1}, k_{-2} = 0.1 s^{-1}, [*S*] = 5 μM, [*I*] = 0 μM, 0.1 μM, 0.2 μM, 0.5 μM, 1 μM, 2 μM, 5 μM, or 10 μM, from top to bottom. B: Time-dependence of the appearance of product P in the absence (black) and in the presence of 5 μM inhibitor (blue), with initial and steady-state reaction velocities, v_0 and v_s, highlighted. C: Initial and steady-state reaction velocities, v_0 (black) and v_s (blue) as a function of the inhibitor concentration. Parameter values are the same as in panel A. D: The inverse of the rate constant *b* (*i.e.* the lifetime) shows a hyperbolic dependence on the inhibitor concentration. The amplitude and curvature depend on K_I, K_S, k_2, and k_{-2}.

We have stated several times before that time-dependent inhibition is observed when k_{-2} is small, and smaller than k_2. Note that k_{-2} can be small, but > 0 s^{-1}, both for non-covalently and for covalently binding inhibitors that interact reversibly with the enzyme. In the extreme case that k_{-2} = 0 s^{-1}, the inhibitor binds irreversibly. Tightly binding enzyme inhibitors are not only useful as potential therapeutics, but also constitute powerful biochemical tools to interrogate enzyme mechanisms (Box 13.5).

BOX 13.5: HIGH-AFFINITY ENZYME INHIBITORS: EXAMPLES AND APPLICATIONS.

High-affinity binding can be achieved by non-covalent and covalent inhibitors that interact with their targets reversibly or irreversibly. Tightly binding inhibitors are classified into sub-groups according to their mechanism of action.

Transition state analogs are compounds that have a chemical structure similar to the transition state (Section 15.2) of the catalyzed reaction. They bind reversibly, but very tightly to the active site of the enzyme, leading to inhibition. If the rate constant of dissociation is very small, binding is essentially irreversible. *Transition state analog inhibitors* are very useful to investigate enzyme mechanisms. The neurotoxin sarin (methylfluorophosphoric acid isopropylester) contains a tetrahedral phosphonate group that mimics the transition state of acetylcholine hydrolysis, and acts as a potent transition state analog inhibitor of acetylcholine esterase.

(*Continued*)

Residues involved in catalysis are typically more reactive than those outside active sites, which makes them prone to reactions with compounds other than the cognate substrate. For example, active-site cysteines form complexes with heavy metal ions, leading to enzyme inactivation. Active-site cysteines also readily react with iodoacetamide, which is often used to count the number of reactive cysteine residues in a protein. A very recent, medicinally relevant addition to the arsenal of cysteine-reactive inhibitors is the compound Paxlovid, a covalent inhibitor that reacts reversibly with the active-site cysteine of the SARS-CoV 2 protease (Figure 13.21).

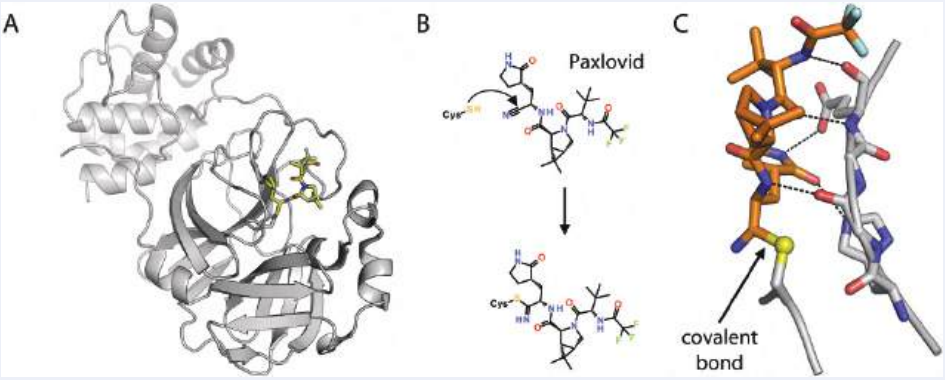


FIGURE 13.21: Covalent inhibition of SARS-CoV 2 protease by Paxlovid. A: Overview of the SARS-CoV 2 protease (PDB-ID 7tll) with the ligand in yellow. B: Reaction of the nitrile group in Paxlovid with cysteine. C: Close-up of the ligand binding site with the ligand in orange and protein residues in gray. Hydrogen bonds between ligand and protein are shown as dashed lines. The covalent bond formed between the nitrile of the inhibitor and the catalytic cysteine residue is marked by an arrow.

Compounds that predominantly react with certain residues are called *group-specific inhibitors*. They can be employed to map active-site residues and serve as covalent inhibitors for disease treatment. Among them are aromatic nitriles such as Paxlovid and acetylenes that react with cysteine and sometimes lysine residues. Epoxides and chloromethyl ketones are often serine-specific.

Molecules that bear similarity to the substrate and cause covalent modification of an enzyme are called *affinity labels*. An example is tosyl-L-phenylalanine chloromethyl ketone that mimics a substrate and binds to the active site of the serine protease chymotrypsin. Once bound the inhibitor covalently modifies a histidine and a serine in the active site.

Suicide inhibitors are molecules that bind to the active site of the enzyme and undergo turnover up to a certain step, but are not converted to product. They remain covalently bound to the enzyme, inhibiting its activity. The antibiotic penicillin is an example of a suicide inhibitor. Penicillin is a β-lactam that binds to the active site of D-alanine transpeptidase, an enzyme involved in cell wall synthesis of bacteria (and not present in humans). The transpeptidase catalyzes the formation of crosslinks between D-alanine and glycine residues in the peptidoglycans of the bacterial cell wall. Along the reaction pathway is a transient acyl intermediate between a serine in the active site and D-alanine. The acyl-enzyme intermediate then reacts with a terminal glycine to form the crosslinking peptide bond. Penicillin binds the active site and forms a stable acyl intermediate with the serine in the active site that cannot undergo turnover, leading to suicide inhibition. Suicide inhibitors hijack the enzyme mechanism, which makes them very specific. Such

(*Continued*)

mechanism-based inhibitors are therefore attractive for therapeutic applications. One such example is suicide inhibitors of the enzyme monoamine oxidase (MAO) for the therapy of depression and Parkinson's disease. These diseases are characterized by low levels of serotonin and dopamine, respectively. MAO catalyzes the deamination and inactivation of both neurotransmitters, reducing their levels in the brain. Thus, specific MAO inhibition by suicide inhibitors offers a treatment.

QUESTIONS

13.1 The derivation of the Michaelis-Menten velocity equation is based on the assumption that $[S] \approx S_0$, which requires $[S] \ll K_S$. Derive the velocity equation for $[S] \approx K_S$.

13.2 What is the velocity v relative to v_{max} at $[S] = 2 \cdot K_S$, $3 \cdot K_S$, $5 \cdot K_S$, and $9 \cdot K_S$? What is the ratio of substrate concentrations for which $v = 0.9\ v_{max}$ and $v = 0.5 \cdot v_{max}$ ($[S]_{0.9}/[S]_{0.5}$)?

13.3 Ethanol (EtOH) is oxidized to acetaldehyde by alcohol dehydrogenase (ADH), using NAD^+ as an oxidant. Starting from a concentration of $[EtOH] = 1.0\ g\ L^{-1}$ in the body, ADH reduces the ethanol concentration by 10% in 1 h. ADH is saturated with EtOH throughout the reaction, and NAD^+ is maintained constant. What is the reaction velocity in $\mu M\ s^{-1}$? $M(EtOH) = 40.07\ g\ mol^{-1}$. What are the molar concentrations and the blood alcohol levels in ‰ at the beginning and at the end, assuming equal distribution throughout the body? Is this assumption valid? How long does it take until the blood alcohol level is below 0.3‰?

13.4 The radical reaction of bromine and hydrogen to give HBr occurs according to the following reaction scheme:

$$Br_2 \xrightarrow{k_1} 2\ Br\bullet$$

$$H_2 + Br\bullet \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} HBr + H\bullet$$

$$Br_2 + H\bullet \xrightarrow{k_3} HBr + Br\bullet$$

$$2Br\bullet \xrightarrow{k_{-1}} Br_2$$

Derive the rate law for HBr production under steady-state assumptions for H• and Br• radicals. Why is this assumption appropriate?

13.5 The enzymatic activity of an enzyme with $K_S = 2$ mM that converts substrate S into product P is measured at an initial substrate concentration S_0 of 10 µM. After 5 min, the substrate concentration is halved. What is the rate constant k, the maximal velocity v_{max}, and the concentration of product after 12 min?

13.6 An enzyme has a histidine residue in the active site ($pK_{A,E} = 6.5$) that must be protonated for catalysis. Its substrate is the deprotonated form of a weak acid with $pK_{A,S} = 4.5$. What is the pH optimum of the enzyme-catalyzed reaction?

13.7 An enzyme has an aspartate in the active site that needs to be deprotonated for substrate binding and catalysis. The dissociation constant for the aspartate is K_A, the dissociation constant of the enzyme-substrate complex is K_S. The rate constant of product formation is k_p. Derive the equation for the reaction velocity. What is the effect of H^+?

13.8 Calculate the inflection point for the following function, expressed as pH

$$K_{app} = K \cdot \frac{1 + \frac{[H^+]}{K_{A'}}}{1 + \frac{[H^+]}{K_A}}$$

13.9 Calculate the ratio of substrate concentrations at which $v = 0.9 \cdot v_{max}$ and $v = 0.1 \cdot v_{max}$ for an enzyme with the apparent Hill coefficient $n_{app} = 2.6$ and without cooperativity.

13.10 An enzyme shows a sigmoidal dependence of the reaction velocity on substrate concentration. The $[S]_{0.9}/[S]_{0.1}$ ratio is 7.6. What is n_{app}? What is the number of substrate binding sites?

13.11 What is the relative distribution of the different enzyme forms and the reaction velocity for an enzyme with two active sites and a cooperativity factor $\alpha = 0.2$ at $[S] = 0.4 \cdot K_S$?

13.12 A dimeric enzyme shows positive cooperativity for substrate binding. The enzyme follows the Monod-Wyman-Changeux model: either both protomers are in the T state that does not bind substrate or both protomers are in the R state that binds substrates with high affinity (dissociation constant K_S). The equilibrium constant for the conformational equilibrium is $L = [T]/[R]$. The rate of product formation is k_p (for each active site). Derive the velocity equation for this enzyme, and generalize the solution to n sites.

13.13 Hexokinase (HK) binds glucose and catalyzes its phosphorylation to glucose-6-phosphate. Fructose binds to the enzyme and inhibits glucose phosphorylation. The dissociation constant of the HK/glucose complex is K_S, the dissociation constant of the HK/fructose complex is K_I. The rate-limiting step of product formation occurs with k_p. Derive the rate equation and compare the apparent K_M and the apparent v_{max} to the values in the absence of inhibition. What are the limiting values for $v_{max,app}$ and $K_{M,app}$ for very low and very high fructose concentration? What happens at $[fructose] = K_I$?

13.14 The activity of an enzyme with $K_S = 50\ \mu M$ and $v_{max} = 1\ \mu M\ s^{-1}$ is measured in the presence of 400 μM substrate and 500 μM of an inhibitor with a K_I of 250 μM. What is the degree of inhibition for competitive, non-competitive, and uncompetitive inhibition?

13.15 What is the relative inhibition of an enzyme by a competitive inhibitor at $[S] = K_S$ and $[I] = K_I$?

13.16 Why is it effective to treat methanol or ethylene glycol poisoning with large (sub-lethal) doses of ethanol?

13.17 An inactive enzyme becomes activated by an allosteric activator A. Derive v/v_{max} as a function of the activator and substrate concentration.

13.18 RNase A has two histidine residues in the active site. All species bind substrate, but for catalysis His12 must be deprotonated, and His119 must be protonated. Deprotonation of His12 favors protonation of His119. Write a reaction scheme and derive the velocity equation as a function of the $[H^+]$ concentration.

13.19 Derive the relationship between IC_{50} and K_I for non-competitive inhibition.

13.20 Show that the following two expressions for K_I^* are equivalent.

$$K_I^* = K_I \cdot \frac{k_{-2}}{k_{-2} + k_2}$$

$$K_I^* = \frac{[E][I]}{[EI] + [EI^*]}$$

13.21 We have derived Michaelis-Menten-like equations for the hyperbolic dependence of the reaction velocities v_0 and v_s on the substrate concentration, with the inhibitor concentration $[I]$ and the inhibition constants K_I and K_I^* as parameters:

$$v_0 = v_{max} \cdot \frac{[S]}{K_S\left(1 + \frac{[I]}{K_I}\right) + [S]}$$

$$v_s = v_{max} \cdot \frac{[S]}{K_S\left(1 + \frac{[I]}{K_I^*}\right) + [S]}$$

Rearrange these equations to show that, for constant substrate concentration $[S]$, v_0 and v_s depend hyperbolically on the inhibitor concentration.

13.22 Derive the equation for $1/b$, the inverse of the inactivation rate constant for time-dependent inhibition, as a function of the inhibitor concentration $[I]$, and determine the amplitude and curvature.

REFERENCES

De La Cruz, E. M., Sweeney, H. L. and Ostap, E. M. (2000) ADP inhibition of myosin V ATPase activity. *Biophys J* **79**(3): 1524–1529.
· product inhibition of the ATPase myosin by ADP

Koshland, D. E. Jr., Nemethy, G. and Filmer, D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5**(1): 365–385.
· description of the Koshland-Nemethy-Filmer model of cooperativity

Monod, J., Wyman, J. and Changeux, J. P. (1965) On the nature of allosteric transitions: A plausible model. *J Mol Biol* **12**: 88–118.
· Monod-Wyman-Changeux model of cooperativity

Morrison, J. F. and Walsh, C. T. (1988) The behavior and significance of slow-binding enzyme inhibitors. *Adv Enzymol Relat Areas Mol Biol* **61**: 201–301.
· formal treatment of time-dependent inhibition including a set of examples

Segel, I. H. (1975) *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady State Enzyme Systems*. New York, Wiley.
· a compendium of steady-state enzyme kinetics

14

Complex reaction schemes and their analysis

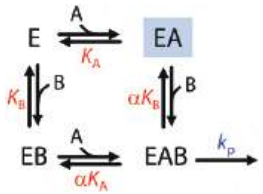
In Chapter 13 we learnt the principles for translating enzymatic reaction schemes into rate equations. This Chapter illustrates that the same basic principles allow us to derive rate equations for binding of two substrates, and presents tools to simplify the analysis of quite complex reaction schemes.

14.1 BINDING OF TWO SUBSTRATES

In sequential mechanisms, two (or more) substrates are bound to the enzyme first, and then react to give the product. Binding can be random, in any possible sequence, or ordered, with a defined sequence of binding events.

14.1.1 RANDOM BINDING

Two substrate molecules can bind to the enzyme in random order, *i.e.* binding of A followed by binding of B or binding of B followed by binding of A leads to the same ternary complex EAB (Scheme 14.1). From EAB, products are formed with the rate constant k_p. The binding equilibria of A and B are described by the equilibrium dissociation constants K_A and K_B. Thermodynamic linkage (Section 2.4) can be accounted for by the factor α (Section 2.4):



scheme 14.1

Product formation from EAB occurs with the reaction velocity

$$v = k_p[EAB] \qquad \text{eq. 14.1}$$

Following the standard procedure (Box 13.2) described several times before, we divide by the total enzyme concentration

$$\frac{v}{[E]_t} = \frac{k_p[EAB]}{[E]+[EA]+[EB]+[EAB]} \qquad \text{eq. 14.2}$$

substitute the concentrations of the enzyme complexes in terms of free ligand concentrations $[A]$ and $[B]$ and their corresponding equilibrium constants, rearrange, and simplify to

$$\frac{v}{v_{max}} = \frac{\frac{[A][B]}{\alpha K_A K_B}}{1+\frac{[A]}{K_A}+\frac{[B]}{K_B}+\frac{[A][B]}{\alpha K_A K_B}} \qquad \text{eq. 14.3}$$

We can now expand eq. 14.3 with $\alpha K_A K_B$ to

$$\frac{v}{v_{max}} = \frac{[A][B]}{\alpha K_A K_B + \alpha K_B[A] + \alpha K_A[B] + [A][B]} \qquad \text{eq. 14.4}$$

Canceling $[B]$ yields

$$\frac{v}{v_{max}} = \frac{[A]}{\alpha K_A\left(1+\frac{K_B}{[B]}\right)+[A]\left(1+\frac{\alpha K_B}{[B]}\right)} \qquad \text{eq. 14.5}$$

By division of the denominator on the right-hand side and multiplying the numerator on the left-hand side with $1+\alpha K_B/[B]$, we can rearrange eq. 14.5 to an equation of the form

$$\frac{v}{v_{max,app}} = \frac{[A]}{K_{M,app}+[A]} \qquad \text{eq. 14.6}$$

with

$$v_{max,app} = \frac{v_{max}}{1+\frac{\alpha K_B}{[B]}} \qquad \text{eq. 14.7}$$

and

$$K_{M,app} = \frac{\alpha K_A\left(1+\frac{K_B}{[B]}\right)}{\left(1+\frac{\alpha K_B}{[B]}\right)} \qquad \text{eq. 14.8}$$

Thus, for a given concentration of A, both $v_{max,app}$ and $K_{M,app}$ depend on the concentration of B. Under the limiting condition that B is saturating, the terms $K_B/[B]$ and $\alpha K_B/[B]$ become small, and both $v_{max,app}$ and $K_{M,app}$ become independent of the concentration of B (Figure 14.1):

$$v_{max,app} \cong v_{max} \qquad \text{eq. 14.9}$$

$$K_{M,app} \cong \alpha K_A \qquad \text{eq. 14.10}$$

In the absence of cooperativity, for $\alpha = 1$, eq. 14.8 simplifies to $K_{M,app} = K_A$, and so $K_{M,app}$ does not depend on $[B]$, but v_{max} still does (Figure 14.1).

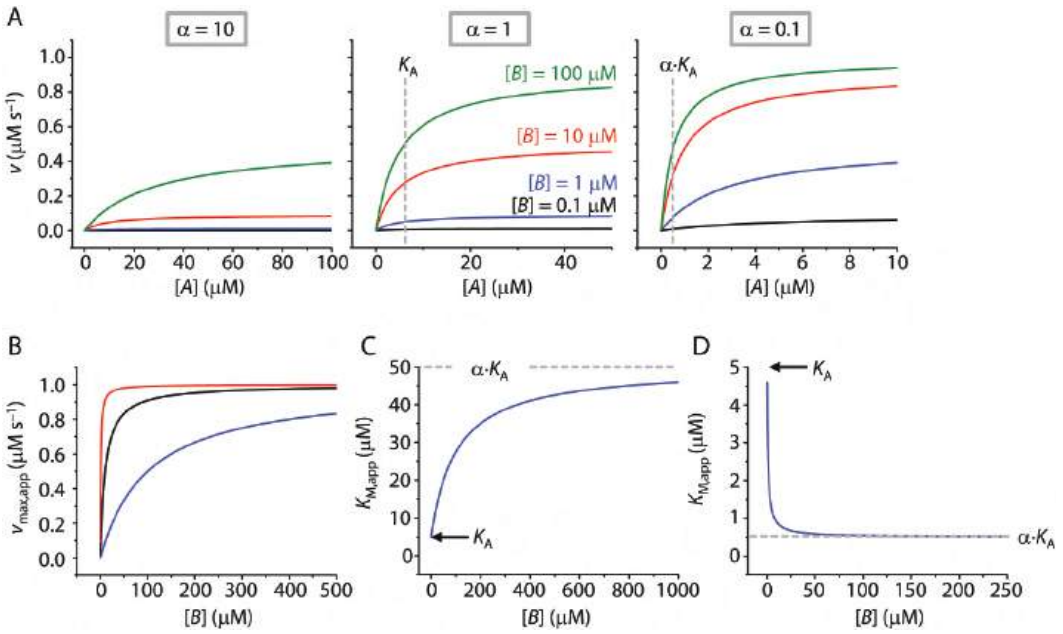


FIGURE 14.1: Random substrate binding. A: Reaction velocity as a function of concentration of substrate A at different concentrations of substrate B in the absence of thermodynamic linkage ($\alpha = 1$), and with positive ($\alpha = 0.1$) and negative ($\alpha = 10$) thermodynamic linkage. $v_{max} = 1\ \mu M\ s^{-1}$, $K_A = 5\ \mu M$, $K_B = 10\ \mu M$, $[B] = 0.1\ \mu M$, 1 μM, 10 μM, 100 μM. B: $v_{max,app}$ as a function of $[B]$ for no thermodynamic linkage ($\alpha = 1$; black), positive ($\alpha = 0.1$; red) and negative thermodynamic linkage ($\alpha = 10$; blue). $v_{max,app}$ reaches v_{max} when $[B]$ is saturating ($[B] > \alpha K_B$). C: $K_{M,app}$ as a function of $[B]$ for $\alpha = 10$. $K_{M,app}$ increases with increasing $[B]$ and approaches αK_A at saturating concentrations of B. D: $K_{M,app}$ as a function of $[B]$ for $\alpha = 0.1$. $K_{M,app}$ decreases with increasing $[B]$ and approaches αK_A at saturating concentrations of B.

Many enzymes form a ternary complex with two ligands to catalyze their reaction. Examples are found among transferases, such as phosphoryl-, acetyl-, or methyl transferases (although some of them follow a ping-pong mechanism (Section 14.2) with a covalently modified enzyme intermediate). Methionine synthase catalyzes methionine synthesis by transfer of a methyl group from methyltetrahydrofolate to homocysteine. Kinetic and equilibrium studies have shown that both substrates are bound in random order, but with high thermodynamic linkage ($\alpha \approx 30$). Other examples are ATP-dependent enzymes that couple ATP binding and hydrolysis to a certain reaction. RNA helicases unwind RNA duplex regions in an ATP-dependent reaction. Some, but not all of them, bind ATP and the RNA duplex in random order, with different levels of thermodynamic linkage.

14.1.2 Ordered binding

In some enzymes, the two substrates have to be bound in a defined order for the reaction to proceed. *Ordered* or *sequential binding* occurs when binding of the first substrate induces a conformational change in the enzyme that leads to formation of the binding site for the second substrate. Substrates have to be bound in a defined order when the binding sites are arranged on the enzyme such that binding of the second substrate blocks the binding site for the first substrate (Figure 14.2). Similarly, release of the products may be ordered.

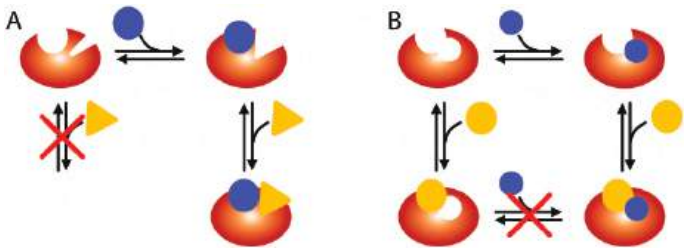
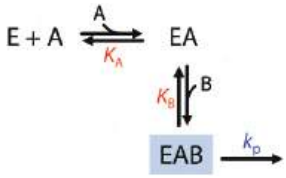


FIGURE 14.2: Sequential binding. A: Binding of substrate 1 (blue) to the enzyme causes a conformatinoal change that leads to formation of the binding site for substrate 2 (orange). B: The binding site for substrate 2 (orange) blocks the binding site for substrate 1 (blue).



scheme 14.2

To evaluate the reaction scheme with ordered binding of substrates A and B (Scheme 14.2), we again begin by formulating the rate of product formation from the EAB complex:

$$v = k_p[EAB] \qquad \text{eq. 14.11}$$

When we now divide by the total enzyme concentration, we only have to take into account free enzyme E, the EA complex, and the EAB complex, as the EB complex does not exist (compare eq. 14.2):

$$\frac{v}{[E]_t} = \frac{k_p[EAB]}{[E]+[EA]+[EAB]} \qquad \text{eq. 14.12}$$

We apply the same sequence of operations as before (Box 13.2): by substituting the concentrations of the enzyme complexes in terms of free ligands and equilibrium constants and division by $[E]$ we obtain

$$\frac{v}{v_{max}} = \frac{\frac{[A][B]}{K_A K_B}}{1+\frac{[A]}{K_A}+\frac{[A][B]}{K_A K_B}} \qquad \text{eq. 14.13}$$

and

$$\frac{v}{v_{max}} = \frac{[A][B]}{K_A K_B + K_B[A]+[A][B]} \qquad \text{eq. 14.14}$$

Canceling $[B]$ gives

$$\frac{v}{v_{max}} = \frac{[A]}{K_A\frac{K_B}{[B]}+[A]\left(1+\frac{K_B}{[B]}\right)} \qquad \text{eq. 14.15}$$

Division by $1+K_B/[B]$ and rearrangement of eq. 14.15 leads to

$$\frac{v}{v_{max,app}} = \frac{[A]}{K_{M,app}+[A]} \qquad \text{eq. 14.6}$$

with

$$v_{max,app} = \frac{v_{max}}{\left(1+\frac{K_B}{[B]}\right)} \qquad \text{eq. 14.16}$$

and

$$K_{M,app} = \frac{K_A K_B}{[B] + K_B} \qquad \text{eq. 14.17}$$

Thus, for ordered binding, both v_{max} and $K_{M,app}$ depend on the concentration of B. If [*B*] is saturating, *i.e.* $[B] \gg K_B$, then $v_{max,app}$ approaches v_{max} and becomes independent of [*B*], and $K_{M,app}$ becomes $K_A K_B/[B]$ (Figure 14.3).

The expression for $v_{max,app}$ in eq. 14.16 is the same expression we have derived for the reaction with random substrate binding (eq. 14.7), except that the factor α for thermodynamic coupling is missing here. In fact, the cooperativity parameter is implicitly contained in the value of K_B. The constant K_B in Scheme 14.2 denotes the equilibrium constant for the dissociation of the EAB complex to give EA and B. No binding of B in the absence of A is equivalent to an infinitely high K_d value of the EB complex. K_B in Scheme 14.2 is much lower than this K_d, and reflects αK_B in Scheme 14.1 (random binding).

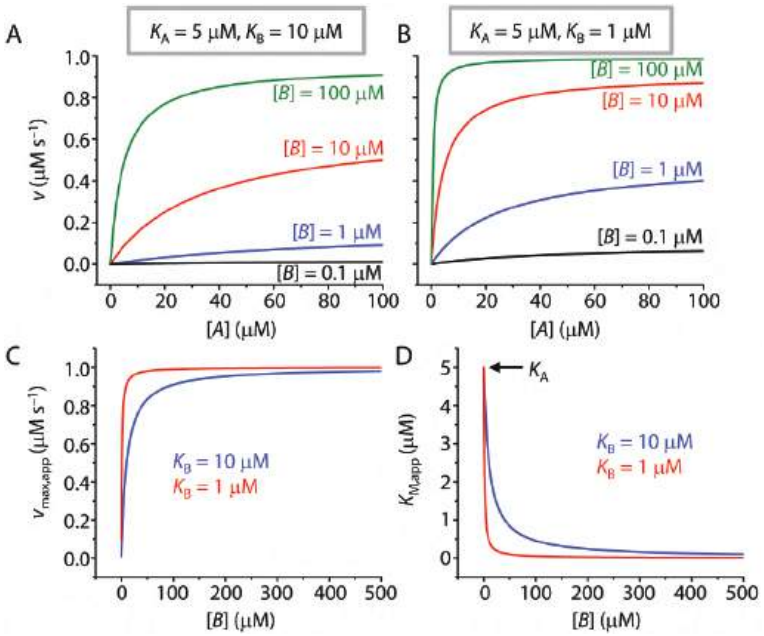
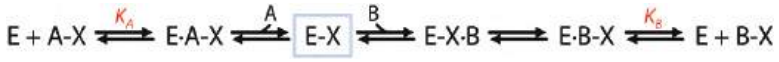


FIGURE 14.3: Ordered sequential binding. A, B: Reaction velocity as a function of concentration of substrate A at different concentrations of substrate B. [*B*] = 1 µM, 10 µM, 100 µM, v_{max} = 1 µM s^{-1}. A: K_A = 5 µM, K_B = 10 µM, B: K_A = 5 µM, K_B = 1 µM. C: $v_{max,app}$ as a function of [*B*] for K_A = 5 µM, K_B = 10 µM (blue) and K_A = 5 µM, K_B = 1 µM (red). For saturating [*B*], $v_{max,app}$ approaches v_{max}. D: $K_{M,app}$ as a function of [*B*] for K_A = 5 µM, K_B = 10 µM (blue) and K_A = 5 µM, K_B = 1 µM (red). For saturating [*B*], $K_{M,app}$ approaches $K_A \cdot K_B/[B]$.

An example of an enzyme with ordered binding of its substrates is serotonin-N-acetyltransferase, an enzyme that catalyzes the transfer of an acetyl group from acetyl-CoA to the terminal amino group of the neurotransmitter serotonin. N-acetyl serotonin is a precursor in the biosynthesis of melatonin, a hormone that regulates the day-night rhythm in humans. Structural studies have shown that binding of acetyl-CoA to the transferase leads to a conformational change of the enzyme and formation of the serotonin binding site. Aspartate transcarbamoylase catalyzes the transfer of a carbamoyl group from carbamoyl phosphate to aspartate, which is converted to dihydroorotate and then orotate, a precursor in pyrimidine biosynthesis. Aspartate transcarbamoylase binds its two substrates in a defined order, first carbamoyl phosphate, and subsequently aspartate. Similar to serotonin-N-acetyltransferase, carbamoyl phosphate binding causes a conformational change of the enzyme that leads to formation of the aspartate binding site. The dissociation of products is also ordered: N-carbamoyl aspartate leaves the enzyme before phosphate.

14.2 PING-PONG MECHANISM

In a *ping-pong mechanism*, substrate A-X binds and transfers a group X to the enzyme. The product leaves the enzyme before substrate B binds to the modified enzyme E-X, and receives group X (Scheme 14.3). Examples of these enzymes are aminotransferases, also called transaminases, and biotin-dependent carboxylases.



scheme 14.3

From initial velocities in the absence of product B-X, we can derive expressions for the apparent $v_{max,app}$ and $K_{M,app}$:

$$v_{max,app} = \frac{v_{max}}{\left(1 + \frac{K_B}{[B]}\right)} \qquad \text{eq. 14.18}$$

and

$$K_{M,app} = \frac{K_A}{\left(1 + \frac{K_B}{[B]}\right)} \qquad \text{eq. 14.19}$$

The direct analysis of this reaction scheme is rather involved, which is why we do not treat it here. We can more easily gain insight into the overall behavior of such an enzyme using the concept of transit times and net rate constants (Section 14.3). From eq. 14.18 and eq. 14.19, we see that at high concentrations of B, $v_{max,app}$ approaches v_{max}, and $K_{M,app}$ approaches K_A (Figure 14.4).

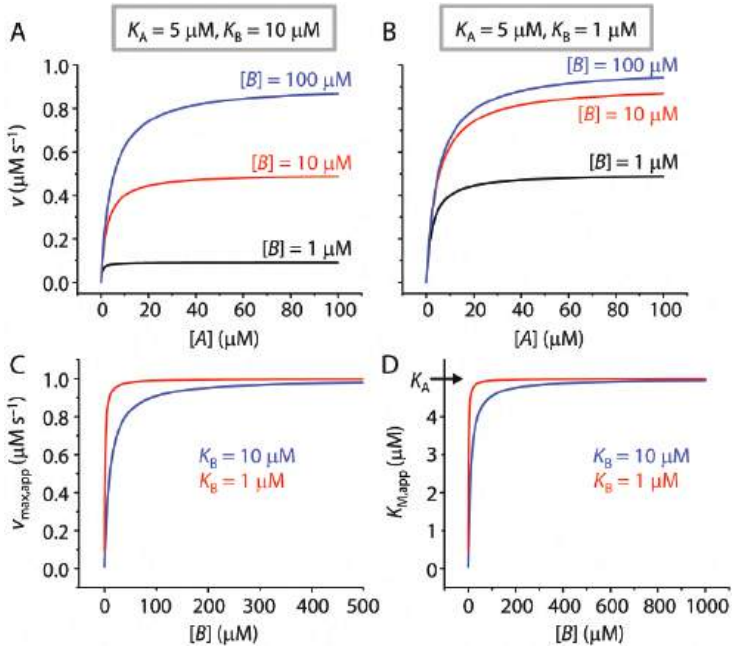


FIGURE 14.4: Ping-pong mechanism. A, B: Reaction velocity as a function of concentration of substrate A at different concentrations of substrate B. [*B*] = 1 µM (black), 10 µM (blue), 100 µM (red). v_{max} is 1 µM s^{-1}. A: K_A = 5 µM, K_B = 10 µM. B: K_A = 5 µM, K_B = 1 µM. C: $v_{max,app}$ as a function of [*B*] for K_A = 5 µM, K_B = 10 µM (blue) and K_A = 5 µM, K_B = 1 µM (red). At saturating concentrations of B, $v_{max,app}$ approaches v_{max}. D: $K_{M,app}$ as a function of [*B*] for K_A = 5 µM, K_B = 10 µM (blue) and K_A = 5 µM, K_B = 1 µM (red). At saturating concentrations of B, $K_{M,app}$ approaches K_A.

An example of an enzyme that follows such a ping-pong mechanism is pyruvate carboxylase, which generates oxaloacetate from pyruvate and CO_2. CO_2 is activated by phosphorylation to carboxyphosphate, $HOCO_2\text{-}PO_3^{2-}$, using ATP as a phosphoryl donor. Pyruvate carboxylase then first binds carboxyphosphate (A-X in Scheme 14.3). Carboxylation of the biotin cofactor leads to carboxybiotin (E-X) and release of phosphate (A). The carboxybiotin cofactor is attached to a lever arm that allows it to swing over a large distance to the bound pyruvate (B), and the CO_2 is transferred from carboxybiotin to pyruvate (corresponding to B-X in Scheme 14.3) to give oxaloacetate (Q). Other enzymes that form an enzyme intermediate are various transferases such as phospho- or acetyltransferases (X = transferred group), oxidoreductases such as flavin reductase (X = hydride), and serine proteases that form an acyl enzyme intermediate (X = acyl peptide, transferred to H_2O).

14.3 NET RATE CONSTANTS AND TRANSIT TIMES

Derivation of the rate equation becomes involved for more complex reaction schemes. Sometimes, insight into the behavior of complex reactions can already be gained by applying some principal concepts, and it may not be necessary to derive the full rate equation. We first consider a reaction scheme of two sequential reactions, a reversible reaction with rate constants k_1 and k_{-1}, followed by an irreversible step with k_2 (Scheme 9.1):

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B \xrightarrow{k_2} C$$

scheme 9.1

How can we intuitively describe the reaction velocity for the conversion of A to C? The overall reaction rate depends on the rate of conversion of A to B, multiplied by the probability that B is converted to C instead of being reconverted to A in the reverse reaction. The rate for the conversion of A to B is $k_1[A]$, and the probability of conversion of B to C instead of back to A is the result of a kinetic partitioning between the forward reaction to C with k_2 and the reverse reaction to A with k_{-1}. The partitioning is described by $k_2/(k_{-1}+k_2)$.

Thus, the *net velocity* v_{net} for the reaction sequence from A to B is the product of these two expressions,

$$v_{net} = \frac{k_1 k_2}{k_{-1} + k_2}[A] \qquad \text{eq. 14.20}$$

and the *net rate constant* k_1' for the reaction A→B is

$$k_1' = \frac{k_1 k_2}{k_{-1} + k_2} \qquad \text{eq. 14.21}$$

To include the conversion from B to C in our consideration, we first have to introduce the concept of *transit times*. (First-order) rate constants have units of s^{-1}, and the inverse of the rate constant thus defines a time $t_{transit}$ that is needed for the corresponding step of the reaction:

$$t_{transit} = \frac{1}{k} \qquad \text{eq. 14.22}$$

This transit time is equal to the time constant τ of the reaction step or the lifetime τ of the reacting species in this step (see Chapter 7). For a four-step reaction sequence from A through B, C, and D to E, with rate constants of k_1, k_2, k_3, and k_4 (Scheme 14.4),

$$A \xrightarrow{k_1} B \xrightarrow{k_2} C \xrightarrow{k_3} D \xrightarrow{k_4} E$$

scheme 14.4

the overall transit time is the sum of the transit times of each individual step. We can therefore sum the individual transit times to calculate the total transit time from A to E. The overall rate constant k is obtained by inverting the total transit time:

$$t_{\text{transit,A}\to\text{E}} = \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \frac{1}{k_4} + \frac{1}{k_5} = \frac{1}{k} \qquad \text{eq. 14.23}$$

We can now combine the two concepts to finish the evaluation of Scheme 14.4, and can calculate the total transit time for the reaction from A to C from the sum of the inversed net rate constants k_1' (eq. 14.21) and k_2:

$$t_{\text{transit,A}\to\text{C}} = \frac{k_{-1} + k_2}{k_1 k_2} + \frac{1}{k_2} = \frac{1}{k} \qquad \text{eq. 14.24}$$

For $k_2 \ll k_1$, the second term in eq. 14.24 is the dominant term which determines the transit time and the overall rate constant of the reaction. In other words: the second step is rate-limiting, and k_2 determines the overall rate constant k.

The concept of net rate constants and transit times can be applied to directly examine the enzymatic reaction occurring after Scheme 14.5, without deriving the rate law step by step.

$$\text{E} + \text{S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \text{ES} \xrightarrow{k_2} \text{E} + \text{P} \qquad \text{scheme 14.5}$$

We analyze the reaction scheme by describing the substrate binding equilibrium by a first-order net rate constant k_1'. For the second step, the net rate constant k'_2 equals k_2:

$$\text{E} + \text{S} \xrightarrow{k'_1} \text{ES} \xrightarrow{k'_2} \text{E} + \text{P} \qquad \text{scheme 14.6}$$

According to eq. 14.21, we can calculate the net rate constant for the binding equilibrium as

$$k'_1 = k_1[S] \cdot \frac{k_2}{k_{-1} + k_2} \qquad \text{eq. 14.25}$$

The total transit time is then

$$t_{\text{transit,S}\to\text{P}} = \frac{1}{k'_1} + \frac{1}{k'_2} = \frac{1}{k_1[S]} \cdot \frac{k_{-1} + k_2}{k_2} + \frac{1}{k_2} \qquad \text{eq. 14.26}$$

The overall rate constant k for the conversion of S to P is the inverse of the total transit time (eq. 14.24), which gives

$$k = \frac{1}{\frac{1}{k_1[S]} \cdot \frac{k_{-1} + k_2}{k_2} + \frac{1}{k_2}} = \frac{k_1 k_2 [S]}{k_{-1} + k_2 + k_1[S]} \qquad \text{eq. 14.27}$$

By canceling k_1 on the right, eq. 14.27 can be expressed in the form

$$k = \frac{k_2[S]}{\frac{k_{-1} + k_2}{k_1} + [S]} \qquad \text{eq. 14.28}$$

The rate constant k is the ratio of reaction velocity v and total enzyme concentration $[E]_t$, and we can re-write eq. 14.28 as

$$\frac{v}{[E]_t} = \frac{k_2[S]}{\frac{k_{-1}+k_2}{k_1}+[S]} \qquad \text{eq. 14.29}$$

Eq. 14.29 is equivalent to the Briggs-Haldane equation (eq. 13.20) with $v_{max} = k_2\ [E]_t$ and $K_M = (k_{-1} + k_2/k_1)$, demonstrating the validity of the simplified approach using transit times and net rate constants.

We can apply the concept of transit times to analyze the ping-pong mechanism (Section 14.2; Scheme 14.3). To this end, we rewrite Scheme 14.3, and describe the rapid equilibrium by the rate constants k_1 and k_{-1}, the transfer of group X to the enzyme by k_2, and the transfer to substrate B by $k_3[B]$ (Scheme 14.7).

$$\mathrm{E + A\text{-}X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E\cdot A\text{-}X \xrightarrow{k_2} E\text{-}X \xrightarrow{k_3[B]} E + B\text{-}X} \qquad \text{scheme 14.7}$$

To calculate the overall transit time $t_{transit} = 1/k$, we need to sum the inverse of the rate constants for each step, k'_1, k'_2, and k'_3. These individual net rate constants are

$$k'_3 = k_3[B] \qquad \text{eq. 14.30}$$

$$k'_2 = k_2 \qquad \text{eq. 14.31}$$

and

$$k'_1 = k_1[A] \cdot \frac{k_2}{k_{-1}+k_2} \qquad \text{eq. 14.32}$$

The total transit time for the transfer of the group X from A to B then is

$$t_{transit,A-X\rightarrow B-X} = \frac{1}{k_1[A]} \cdot \frac{k_{-1}+k_2}{k_2} + \frac{1}{k_2} + \frac{1}{k_3[B]} = \frac{1}{k} = \frac{[E]_t}{v} \qquad \text{eq. 14.33}$$

Compared to a simple enzymatic reaction that consists of formation of the enzyme-substrate complex followed by product formation (eq. 14.26), the only difference in eq. 14.33 is the last term that increases the transit time and thus reduces the overall reaction velocity *v*. At high concentrations of B, this term becomes small and can be neglected. The rate of product formation is then limited by the formation of E-X from EA (k_2) that becomes the rate-limiting step (Chapter 10). This is in agreement with the prediction from the explicit equation for $v_{max,app}$ (eq. 14.18, see Figure 14.4): for high concentrations of B, $1+K_B/[B]$ reduces to unity, and $v_{max,app}$ becomes independent of $[B]$ and equals v_{max}. Although we have not explicitly derived the rate equation, we can already make predictions for the behavior of such a reaction from these considerations.

We have seen that the net rate constant for an irreversible step equals the microscopic rate constant k for this step. For bimolecular reaction steps, the net first-order rate constant is the product of the microscopic second-order rate constant k_n and the concentration of the ligand X, *i.e.* $k[X]$. Net rate constants for reversible steps are calculated from the product of the forward reaction and the partitioning ratio for forward and reverse reactions (eq. 14.21). The evaluation of net rate constants is always performed backwards, moving step by step from the end of the reaction scheme to the first step. For the reaction scheme

$$\mathrm{E \underset{k_{-1}}{\overset{k_1[S]}{\rightleftharpoons}} E_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E_2 \xrightarrow{k_3} E} \qquad \text{scheme 14.8}$$

we can formulate the net rate constants k'_1, k'_2, and k'_3, starting with k'_3 as

$$k'_3 = k_3 \qquad \text{eq. 14.34}$$

$$k'_2 = k_2 \cdot \frac{k'_3}{k_{-2} + k'_3} = k_2 \cdot \frac{k_3}{k_{-2} + k_3} \qquad \text{eq. 14.35}$$

and

$$k'_1 = k_1[A] \cdot \frac{k'_2}{k_{-1} + k'_2} = k_1[A] \cdot \frac{k_2 \cdot \dfrac{k_3}{k_{-2} + k_3}}{k_{-1} + k_2 \cdot \dfrac{k_3}{k_{-2} + k_3}} \qquad \text{eq. 14.36}$$

Again, we can calculate the overall net rate constant from the inverse of the sum of the net rate constants for each step (eq. 14.23). The concept of net rate constants and transit times is particularly helpful to analyze complex multi-step reactions.

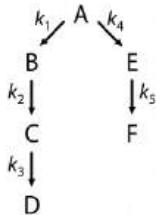
QUESTIONS

14.1 From a thermodynamic perspective, the change in free energy for binding of two substrates to an enzyme must be independent of the order of binding. Why can ordered binding be observed?

14.2 An enzyme follows the ping-pong mechanism. In the first step, E reacts with substrate A to give product P and E'. E' then binds substrate B. Conversion of B to product Q regenerates E. Using the steady-state approximation for E', derive the rate equation for the generation of the products.

14.3 An enzyme exists in two conformations, E_1 and E_2, which are interconverted with the rate constants k_1 and k_{-1}. Only E_2 binds substrate S. The bimolecular rate constant for substrate binding is k_2, the ES complex dissociates to E_2 and S with k_{-2}. Product formation occurs with k_p. Derive the overall rate constant of product formation using the principle of net rate constants and transit times.

14.4 Consider a branched reaction with a sequence of steps in each branch according to the reaction scheme. Determine the ratios of the end products D and F of both paths.



14.5 The branched reaction in Question 14.4 occurs with the following rate constants: $k_1 = 0.5\ s^{-1}$, $k_2 = 0.1\ s^{-1}$, $k_3 = 0.8\ s^{-1}$, $k_4 = 0.6\ s^{-1}$, $k_5 = 0.2\ s^{-1}$. Estimate and calculate the ratio of [*D*] and [*F*] under these conditions.

REFERENCES

Cleland, W. W. (1975) Partition analysis and the concept of net rate constants as tools in enzyme kinetics. *Biochemistry* **14**(14): 3220–3224.
· a clear description of the concept of net rate constants

Hickman, A. B., Namboodiri, M. A., Klein, D. C. and Dyda, F. (1999) The structural basis of ordered substrate binding by serotonin N-acetyltransferase: Enzyme complex at 1.8 A resolution with a bisubstrate analog. *Cell* **97**(3): 361–369.
· structure of serotonin-N-acetyltransferase in complex with a bisubstrate inhibitor that suggests conformational changes upon acetyl-CoA binding

Porter, R. W., Modebe, M. O. and Stark, G. R. (1969) Aspartate transcarbamylase. Kinetic studies of the catalytic subunit. *J Biol Chem* **244**(7): 1846–1859.

Hsuanyu, Y. and Wedler, F. C. (1988) Kinetic mechanism of catalytic subunits (c3) of *E. coli* aspartate transcarbamylase at pH 7.0. *Biochim Biophys Acta* **957**(3): 455–458.
· kinetic studies that demonstrate ordered binding of substrates to aspartate transcarbamoylase

Samatanga, B. and Klostermeier, D. (2014) DEAD-box RNA helicase domains exhibit a continuum between complete functional independence and high thermodynamic coupling in nucleotide and RNA duplex recognition. *Nucleic Acids Res* **42**(16): 10644–10654.
· thermodynamic study that shows different order of binding and different thermodynamic linkage in ATP and RNA binding to RNA helicases

Segel, I. H. (1975) *Enzyme Kinetics: Behavior and Analysis of Rapid Equilibrium and Steady State Enzyme Systems*. New York, Wiley.
· a compendium of steady-state enzyme kinetics

Tanner, J. J., Lei, B., Tu, S. C. and Krause, K. L. (1996) Flavin reductase P: Structure of a dimeric enzyme that reduces flavin. *Biochemistry* **35**(42): 13531–13539.
· structural basis for ping-pong mechanism by flavin reductase

Taurog, R. E., Jakubowski, H. and Matthews, R. G. (2006) Synergistic, random sequential binding of substrates in cobalamin-independent methionine synthase. *Biochemistry* **45**(16): 5083–5091.
· analysis of random substrate binding to methionine synthase

Wang, J., Stieglitz, K. A., Cardia, J. P. and Kantrowitz, E. R. (2005) Structural basis for ordered substrate binding and cooperativity in aspartate transcarbamoylase. *Proc Natl Acad Sci U S A* **102**(25): 8881–8886.
· structural study that shows conformational changes upon aspartate binding to aspartate transcarbamoylase

Wedler, F. C. and Gasser, F. J. (1974) Ordered substrate binding and evidence for a thermally induced change in mechanism for *E. coli* aspartate transcarbamylase. *Arch Biochem Biophys* **163**(1): 57–68.
· kinetic study that demonstrates ordered release of products from aspartate transcarbamoylase

15

Temperature dependence of rate constants

15.1 THE ARRHENIUS EQUATION

Reaction velocities and rate constants depend on temperature. As a rule of thumb, reaction velocities increase by a factor of 2–4 for each increase in temperature by 10°C. Empirically, it has been found that the temperature dependence of rate constants can be described by an exponential law as

$$k(T) = A \cdot e^{-E_a/RT} \qquad \text{eq. 15.1}$$

Eq. 15.1 is called the *Arrhenius equation*. E_a is the activation energy of the reaction, the energy difference between the reactants and an activated complex formed by the reactants that decays to products (see Figure 15.1). The Arrhenius equation tells us that slow reactions, with small rate constants, have a high activation energy, whereas fast reactions, with large rate constants, have a low activation energy.

According to eq. 15.1, reactions inevitably become faster and faster with increasing temperatures. By measuring rate constants at different temperatures, we can determine the activation energy for the reaction using the Arrhenius equation. In the following Sections, we will discuss two different concepts, the *transition state theory* (Section 15.2) and the *collision theory* (Section 15.3), that provide different molecular explanations for the temperature dependence of rate constants described by the Arrhenius equation.

15.2 TRANSITION STATE THEORY

Reactants and products of a reaction have a free energy G, which is the sum of their standard free energies of formation (Section 1.6.3; Figure 1.24). Reactants A and B and products P are separated from each other by an energy barrier, the *activation barrier* (Figure 15.1). The activation barrier reflects the high energy of an activated collision complex $AB^{\neq}$ of the reactants that is ready to

dissociate into the products of the reaction. Parameters pertaining to the activated complex, its formation and dissociation, are typically marked by the superscript ≠. Instead of E_a, we can also write $\Delta G^{\neq}$.

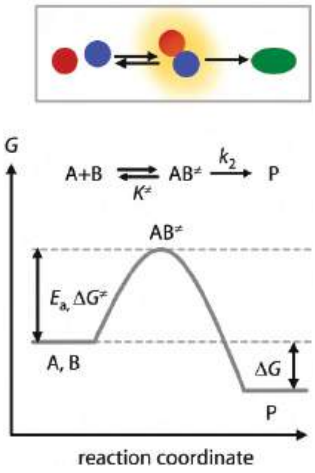


FIGURE 15.1: Transition state theory. The reactants A and B and the products P have different free energies *G*. The difference between their free energies is the change in free energy during the reaction, ΔG. Reactants and products are separated by an activation barrier that represents a high-energy collision complex of A and B that dissociates into the products P. The height of the activation barrier relative to the free energies of A and B is the activation energy E_a or $\Delta G^{\neq}$.

We now assume that the reactants A and B and the activated complex $AB^{\neq}$ are in rapid equilibrium, and the velocity of the dissociation of $AB^{\neq}$ determines the overall reaction velocity of product formation (Scheme 15.1).

$$A + B \overset{K^{\neq}}{\rightleftharpoons} AB^{\neq} \xrightarrow{k_2} \text{products} \qquad \text{scheme 15.1}$$

The reaction velocity then is

$$v = k_2 \left[AB^{\neq} \right] \qquad \text{eq. 15.2}$$

The concentration of the activated complex can be expressed in terms of the equilibrium constant of activation, $K^{\neq}$, as

$$K^{\neq} = \frac{\left[AB^{\neq} \right]}{[A][B]} \qquad \text{eq. 15.3}$$

Note that $K^{\neq}$ is the equilibrium constant of $AB^{\neq}$ complex formation. Eq. 15.3 can be rearranged to give the concentration of the activated complex:

$$\left[AB^{\neq} \right] = K^{\neq} [A][B] \qquad \text{eq. 15.4}$$

We know that equilibrium constants are related to free energy changes under standard conditions. Together with the Gibbs-Helmholtz equation (eq. 1.111), we can therefore write for $K^{\neq}$:

$$K^{\neq} = e^{-\Delta G^{\neq}/RT} = e^{-\Delta H^{\neq}/RT} e^{\Delta S^{\neq}/R} \qquad \text{eq. 15.5}$$

$\Delta G^{\neq}$ is the free energy of activation, $\Delta H^{\neq}$ is the activation enthalpy, and $\Delta S^{\neq}$ is the activation entropy. We can now express the concentration of the activated complex in terms of the activation enthalpy and entropy, and as a function of the concentration of the reactants A and B:

$$\left[AB^{\neq}\right] = e^{-\Delta H^{\neq}/RT} e^{\Delta S^{\neq}/R} [A][B] \qquad \text{eq. 15.6}$$

Now we have to relate k_2 to molecular properties of the transition state. The rate constant k_2 for the dissociation of the transition state to the product(s) will depend on the vibrational energy of the scissile bond, the bond that needs to be broken in the reaction from transition state to products. The vibrational energy depends on the thermal energy of the molecule, which is the product of the Boltzmann constant k_B multiplied by the temperature T (see Section 21.1 and eq. 21.5). The vibration frequency ν can be expressed as

$$\nu = \frac{k_B T}{h} \qquad \text{eq. 15.7}$$

with the Planck constant h. We have come across this vibration frequency before when we discussed the kinetic isotope effect (Chapter 10), which is caused by different vibration frequencies of the scissile bond depending on the masses of the bonded atoms. We can now express the rate constant k_2 in Scheme 15.1 by multiplying this frequency with a transmission coefficient κ.

$$k_2 = \kappa \frac{k_B T}{h} \qquad \text{eq. 15.8}$$

The coefficient κ describes the probability of the bond to dissociate in the direction of the product(s). Substituting eq. 15.6 and eq. 15.8 into eq. 15.2 now yields

$$v = \kappa \frac{k_B T}{h} e^{-\Delta H^{\neq}/RT} e^{\Delta S^{\neq}/R} [A][B] \qquad \text{eq. 15.9}$$

Eq. 15.9 is called the *Eyring equation*. It provides an expression for the reaction velocity v for a bimolecular reaction of A and B in the form of

$$v = k[A][B] \qquad \text{eq. 15.10}$$

with the rate constant k

$$k = A \cdot e^{-\Delta H^{\neq}/RT} \qquad \text{eq. 15.11}$$

and the pre-exponential factor A:

$$A = \kappa \frac{k_B T}{h} e^{\Delta S^{\neq}/R} \qquad \text{eq. 15.12}$$

These equations relate the *activation energy* $\Delta G^{\neq}$ for a reaction to its rate constant k. Eq. 15.12 is formally equivalent to the empirical Arrhenius equation (eq. 15.1), with $E_a = \Delta H^{\neq}$. A high activation enthalpy corresponds to a small k and a slow reaction, and a low activation enthalpy corresponds to a high k and a fast reaction. The pre-exponential factor A in eq. 15.11 contains the entropy of activation $\Delta S^{\neq}$.

Note that the activation energies $\Delta G^{\neq}$ for the forward and reverse reactions differ if reactants and products have different free energies. For a spontaneous reaction with $\Delta G < 0$, $\Delta G_f^{\neq}$ for the forward reaction is smaller than $\Delta G_r^{\neq}$ for the reverse reaction, and the forward reaction will be faster than the reverse reaction. The equilibrium constant K_{kin} for product formation, which is the

ratio of k_f and k_r, is thus larger than unity. Under standard conditions we can also relate the energy difference ΔG^0 to the equilibrium constant K_{eq}, which is larger than unity for exergonic reactions with $\Delta G^0 < 0$. Kinetic and energetic considerations are thus consistent. For an endergonic reaction ($\Delta G^0 > 0$, $K_{eq} < 1$), $\Delta G_f^{\neq}$ is larger than $\Delta G_r^{\neq}$, and k_f is smaller than k_r, consistent with $K_{kin} = k_f/k_r < 1$. For a reaction in equilibrium, we have $\Delta G^0 = 0$ and $K_{eq} = 1$ on the energy side, and $\Delta G_f^{\neq} = \Delta G_r^{\neq}$, $k_f = k_r$, and $K_{kin} = k_f/k_r = 1$ on the kinetic side (Figure 15.2).

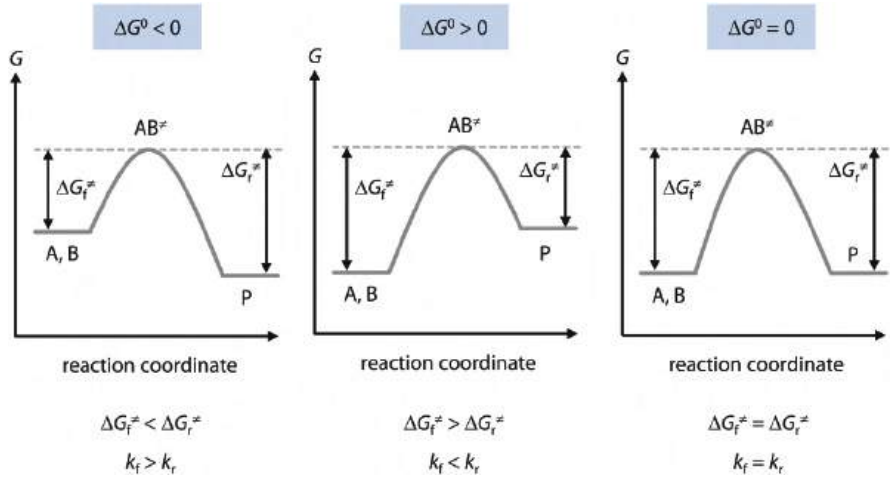


FIGURE 15.2: Activation energy and rate constants for forward and reverse reactions. Energy diagrams for an exergonic (left, $\Delta G^0 < 0$, $K_{eq} > 1$) and an endergonic reaction (center, $\Delta G^0 > 0$, $K_{eq} < 1$) and for a reaction in equilibrium (right, $\Delta G^0 = 0$, $K_{eq} = 1$). The activation energies for the forward and reverse reaction, $\Delta G_f^{\neq}$ and $\Delta G_r^{\neq}$, are indicated. For the exergonic reaction, k_f exceeds k_r, and the equilibrium constant K_{kin} calculated from the rate constants is larger than unity, in agreement with the K_{eq} from ΔG^0. For the endergonic reaction, k_f is smaller than k_r, giving $K_{kin} < 1$. In equilibrium, $\Delta G_f^{\neq}$ equals $\Delta G_r^{\neq}$, k_f equals k_r, and K_{kin} equals 1. In all three scenarios, the energetic and kinetic considerations are consistent.

15.3 COLLISION THEORY

An alternative approach to reaction velocities assumes that the reaction velocity is proportional to the number of collisions Z of the reactants, and to the probability that they collide with sufficient energy for the reaction to occur. The kinetic energy required for a productive collision is called the activation energy, E_a. The probability that the kinetic energy of the colliding molecules is equal to the activation energy is $e^{-E_a/RT}$. Multiplication of these two expressions yields the reaction velocity v:

$$v = Z \cdot e^{-E_a/RT} \qquad \text{eq. 15.13}$$

The number of collisions Z will depend on the concentrations of the reactants, and we can express Z as

$$Z = X \cdot [A][B] \qquad \text{eq. 15.14}$$

Yielding

$$v = X \cdot [A][B] \cdot e^{-E_a/RT} \qquad \text{eq. 15.15}$$

Again, this is an expression for the reaction velocity v in the form of

$$v = k[A][B] \qquad \text{eq. 15.16}$$

this time with

$$k = X \cdot e^{-E_a/RT} \qquad \text{eq. 15.17}$$

Eq. 15.17 has the same form as the empirical Arrhenius equation (eq. 15.1), with $A = X$. The pre-exponential factor X is a probability factor that takes into account speed, size, and orientation of reactants. The units depend on the type of the reaction: for a first-order reaction the units are s^{-1}, for a second-order reaction $M^{-1}\ s^{-1}$. E_a is the activation free energy of the reaction ($\Delta G^{\neq}$). In this case, slow reactions are associated with a high activation free energy, fast reactions with a low activation free energy.

Although eq. 15.9 (transition state theory) and eq. 15.15 (collision theory) have the same form, the molecular meanings of the pre-exponential factors and activation energies are different. Both theories explain the concentration and temperature dependence of rate constants, but they describe the effects with different parameters. The increase in rate constants with temperature is interpreted as accelerated breaking of the relevant bond in the transition state because of a higher vibration frequency (transition state theory) or as increased collision frequency due to higher kinetic energy of the reactants (collision theory). The increase in reaction velocities with increased reactant concentration is caused by a shift in the activation equilibrium to the transition state (transition state theory) or a higher frequency of collisions with less distance separating the reactant molecules (collision state theory).

15.4 KINETIC AND THERMODYNAMIC CONTROL OF REACTIONS

Thermodynamic considerations allow us to predict if a reaction occurs spontaneously (Section 1.6), but do not provide any information on the rate with which the reaction occurs. The velocity of a reaction depends on the height of the energy barrier that separates the reactant(s) from the product(s). Therefore, a thermodynamically favorable reaction might occur very slowly if the activation energy is high. Compounds that are stable because they react slowly, although the reaction would be energetically downhill, are called *metastable*. One prominent example of a metastable compound is ATP. We have discussed ATP hydrolysis as a highly exergonic reaction several times. Despite the large negative ΔG^0, ATP hydrolysis is rather slow because of a high activation barrier for hydrolysis, and ATP is quite stable. The situation is different *in vivo*, where ATP hydrolysis is catalyzed by enzymes that reduce the activation barrier and increase the rate of ATP hydrolysis by several orders of magnitude (Section 16). Peptides are another biochemical example of metastable compounds. Peptide bond formation is an endergonic reaction that is coupled to ATP hydrolysis in the cell. The inverse reaction, peptide bond hydrolysis, is exergonic, but the activation energy is very high, and the spontaneous hydrolysis of peptides is hardly observed.

It is important to appreciate the implications of the interplay between kinetics and thermodynamics for parallel reactions (Section 8.2). If a reactant A can be converted to B or C, and the free energy change ΔG_r is more negative for the reaction to C than for the reaction to B (Figure 15.3), then C will be the more populated (or even the exclusive) reaction product at the end of the reaction: the reaction is under *thermodynamic control*. However, the time that is necessary until all three compounds have equilibrated depends on the rate constants for individual steps, and these depend on the activation energies. If the activation barrier between A and the stable product C is much higher than the energy barrier that separates A from the less stable product B (Figure 15.3), then A will be converted much more rapidly to B than to C, and B is generated at the beginning of the reaction. The (transient) accumulation of B is a *kinetically controlled*

reaction. During the slow equilibration, B can then be re-converted to A and either revert to B or slowly react to C. Once C is generated, the reaction back to A is slow because of the high energy barrier. Nevertheless, after some time, the thermodynamically controlled equilibrium fractions of A, B, and C are reached. In extreme cases, depending on the differences in ΔG^0 and $\Delta G^{\neq}$ for the two branches of the reaction, the less stable products might be exclusively generated in a kinetically controlled reaction, followed by an infinitely slow equilibration to the thermodynamically controlled product. By performing a reaction under kinetic control, thermodynamically disfavored reaction products can be captured.

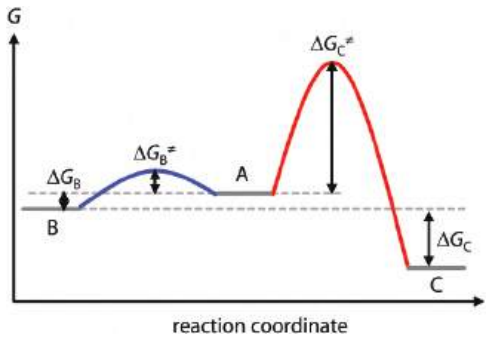


FIGURE 15.3: Thermodynamic and kinetic control. Reactant A can react to products B or C (parallel reaction). The reaction to B has a low activation energy and is rapid. The reaction to C has a high activation energy and is slow. C is the more stable product. Under kinetic control, B will be formed. Under thermodynamic control, C is the dominant product.

QUESTIONS

15.1 The velocity of a reaction increases two-fold when the temperature is increased from 25°C to 37°C. Calculate the activation energy. How does the activation energy change when the increase in the rate constant k is four-fold?

15.2 Van't Hoff's rule (not to be confused with the van't Hoff equation) states that an increase in temperature by 10 K leads to a 2–4-fold increase in reaction velocity. What is the corresponding activation energy interval for a starting temperature T_1 = 298 K? We assume that the pre-exponential factor and the activation energy in the Arrhenius equation are independent of temperature over the temperature interval considered.

15.3 You think about storing a protein solution at 4°C instead of on ice. The protein degradation reaction has an activation energy of E_A = 50 kJ·mol^{-1}. What factor $n = k_2/k_1$ relates the rate constants at these temperatures?

15.4 A protein folds to the native state that is 30 kJ mol^{-1} more stable than the unfolded state. The activation barrier for folding is E_a = 40 kJ mol^{-1}. The protein can also form a non-functional intermediate (ΔG_r = –20 kJ mol^{-1}). The activation energy is E_a = 5 kJ mol^{-1}. What are the (relative) rate constants for the reaction from U to I, I to U, U to N, and N to U? What are the approximate fractions of I and N under kinetic control, what is the fraction of N under thermodynamic control? Which rate constant determines the time scale of folding into the native state? Assume in a first approximation that the pre-exponential factor of the Arrhenius equation is identical for all four reactions. What are the equilibrium constants and the fractions of N, U, and I in equilibrium?

15.5 DNA strands (A,B) associate rapidly to mismatched intermediates that slowly rearrange to the correctly base-paired duplex P. Derive an expression for the activation energy of the overall reaction from A and B to P. What is the reason for a "negative activation energy"?

15.6 Bimolecular reactions can be described by a first step that is the encounter of the reactants, and a second step that is the chemical reaction to products. Derive the overall rate constant and simplify for a diffusion-controlled reaction in which the formation of the encounter complex (diffusion) is rate-limiting, and the reaction to products is rapid, and for an activation-controlled reaction where the reaction to products is rate-limiting.

16

Principles of catalysis

Catalysts accelerate chemical reactions without affecting the equilibrium constant. From the relationship between forward and reverse rate constants and the equilibrium constant (eq. 8.21, eq. 8.23), we can infer that catalysis must accelerate forward and reverse reactions by the same factor. According to transition state theory (Section 15.2), catalysts accelerate chemical reactions by stabilization of the active complex, the transition state (Figure 16.1). As a consequence, the activation barrier is reduced, and the rate constant is increased. By definition, a catalyst is regenerated at the end of the reaction, meaning it is not consumed but can catalyze multiple turnovers.

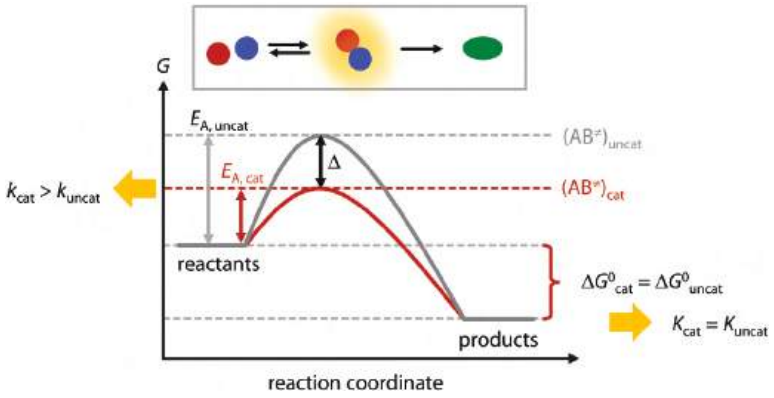


FIGURE 16.1: Catalysis and the transition state. A catalyst stabilizes the transition state $AB^{\neq}$ of a reaction. The concomitant reduction of the energy barriers for forward and reverse reactions (Δ) leads to an acceleration ($k_{cat} > k_{uncat}$) of the reaction in both directions. The free energies of the reactants and products are unaltered, and the equilibrium constant K of the reaction is therefore not changed.

In *heterogeneous catalysis*, the catalyst and the substrates and products of the reaction catalyzed are present in different phases. Heterogeneous catalysis is frequently encountered in technical applications, such as the reaction of gases on the surface of a solid metal catalyst. In *homogeneous catalysis*, the catalyst and the substrates and products are in the same phase. This is the case for enzyme-catalyzed reactions. Heterogeneous and homogeneous catalysis are formally equivalent, and

lead to analogous expressions for the rate laws. We will first derive rate equations for heterogeneous, surface-catalyzed reactions (Section 16.1), and will see how the same concepts underlie homogeneous catalysis by enzymes (Section 16.2). In Sections 16.2.1–16.2.3, the main catalytic principles will be highlighted.

16.1 HETEROGENEOUS CATALYSIS: THE LANGMUIR-HINSHELWOOD FORMALISM FOR SURFACE-CATALYZED REACTIONS

Solid-phase catalysis is a prime example for heterogeneous catalysis. The reactants, in the gas or liquid phase, bind to the surface of the solid catalyst, often metal, and the reaction takes place on the surface. Similar to enzyme-catalyzed reactions (see Sections 3.4, 13.4, and 13.5), the reaction velocity of such a reaction depends on the degree of saturation of the binding sites available on the catalyst. For enzyme-catalyzed reactions (eq. 13.8), we have derived the velocity equations by starting from the relationship between the velocity of the reaction catalyzed and the concentration of the enzyme-substrate complex (eq. 13.3). We can use the same concept to derive the velocity equation for the conversion of the substrate S to the product P by heterogeneous catalysis: the reaction velocity v is proportional to the fraction θ of binding sites on the surface of the catalyst that is occupied with substrate:

$$v = k \cdot \theta \qquad \text{eq. 16.1}$$

In analogy to eq. 13.3, the proportionality constant is the rate constant k of the reaction from S to P on the surface. The fraction θ can be calculated from the equilibrium of adsorption and desorption of the substrate to and from the surface (Figure 16.2).

The adsorption velocity v_{ads} is proportional to the concentration of the substrate, which we express as its pressure p in the gas phase, and to the number of binding sites available on the surface.

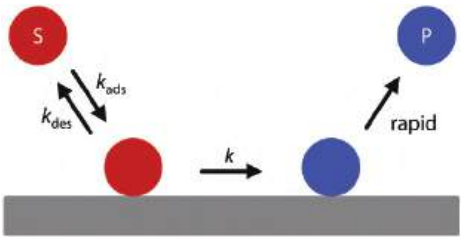


FIGURE 16.2: Heterogeneous catalysis. A substrate S (red) from the gas (or liquid) phase is in equilibrium with binding sites on the surface of a catalyst (gray). On the surface, it is converted into the product P (blue), which rapidly desorbs from the catalyst. k_{ads}, k_{des}: rate constants for adsorption and desorption of the substrate, k: rate constant of substrate conversion into product on the surface.

The number of binding sites available for adsorption, in turn, is the product of the total number of binding sites n_{tot}, and the fraction of sites that are unoccupied (1–θ), giving:

$$v_{ads} = k_{ads} \cdot p \cdot (1 - \theta) \cdot n_{tot} \qquad \text{eq. 16.2}$$

By analogy, the desorption, *i.e.* the dissociation of substrate from the surface, occurs with the velocity v_{des} according to

$$v_{des} = k_{des} \cdot \theta \cdot n_{tot} \qquad \text{eq. 16.3}$$

In equilibrium, v_{ads} equals v_{des}, and by equating eq. 16.2 and eq. 16.3, we obtain

$$k_{ads} \cdot p \cdot (1-\theta) \cdot n_{tot} = k_{des} \cdot \theta \cdot n_{tot} \qquad \text{eq. 16.4}$$

which we can rearrange to

$$\frac{1-\theta}{\theta} = \frac{k_{des} \cdot n_{tot}}{k_{ads} \cdot p \cdot n_{tot}} \qquad \text{eq. 16.5}$$

or

$$\frac{1}{\theta} - 1 = \frac{k_{des}}{k_{ads} \cdot p} \qquad \text{eq. 16.6}$$

which equals

$$\frac{1}{\theta} = \frac{k_{des}}{k_{ads} \cdot p} + 1 \qquad \text{eq. 16.7}$$

Using the same denominator for the right-hand side of eq. 16.7 gives

$$\frac{1}{\theta} = \frac{k_{des} + k_{ads} \cdot p}{k_{ads} \cdot p} \qquad \text{eq. 16.8}$$

which we can invert to

$$\theta = \frac{k_{ads} \cdot p}{k_{des} + k_{ads} \cdot p} \qquad \text{eq. 16.9}$$

By dividing all terms on right-hand side by k_{des}, we obtain

$$\theta = \frac{\frac{k_{ads}}{k_{des}} \cdot p}{1 + \frac{k_{ads}}{k_{des}} \cdot p} \qquad \text{eq. 16.10}$$

We now substitute b for the ratio of the rate constants:

$$b = \frac{k_{ads}}{k_{des}} \qquad \text{eq. 16.11}$$

which gives

$$\theta = \frac{b \cdot p}{1 + b \cdot p} \qquad \text{eq. 16.12}$$

Eq. 16.12 describes the dependence of the fraction of binding sites occupied, θ, on the pressure p of the substrate S in the gas phase. This expression is known as the *Langmuir binding isotherm*. By inserting this expression into eq. 16.1, we obtain the velocity v of the reaction on the surface as

$$v = k \cdot \frac{b \cdot p}{1 + b \cdot p} \qquad \text{eq. 16.13}$$

We see from eq. 16.13 that the reaction velocity for a surface-catalyzed reaction shows a hyperbolic dependence on the concentration of the reactant, expressed in terms of the pressure p in eq. 16.13, the same way as the velocity of enzymatically catalyzed reactions shows a hyperbolic dependence on the concentration [S] of the substrate (eq. 13.8, eq. 13.9). Similar to the considerations we made

before for low and high substrate concentrations of the enzyme-catalyzed reaction, the reaction velocity is proportional to p at low pressure ($1 + bp \rightarrow 1$ for $p \rightarrow 0$), meaning the surface-catalyzed reaction follows a first-order rate law. At high pressure, the reaction becomes independent of p and follows a zero-order rate law (the denominator can be approximated by bp). The formalism described here is termed *Langmuir-Hinshelwood formalism* for a unimolecular, surface-catalyzed reactions.

16.2 HOMOGENEOUS CATALYSIS BY ENZYMES

In homogeneous catalysis by enzymes, the enzyme corresponds to the surface catalyst, the substrate corresponds to the reactants provided in the gas phase. Catalysis, achieved by adsorption of the reactants to the surface of the catalyst in heterogeneous catalysis, is achieved by binding of the substrate to the enzyme surface. The analogy extends to the velocity equations: eq. 16.13 describing the hyperbolic dependence of the reaction velocity of a heterogeneously catalyzed reaction is equivalent to the hyperbolic dependence of the reaction velocity of a homogeneously catalyzed enzymatic reaction (eq. 13.9). Comparing eq. 16.13 and eq. 13.9 tells us that K_S in homogeneous catalysis by enzymes corresponds to $1/b$ in heterogeneous catalysis. This is commensurate with the definition of b as the ratio of the rate constants for adsorption and desorption, k_{ads} and k_{des}, which defines b as an association constant (eq. 16.11). In contrast, K_S is a dissociation constant, and hence the inverse.

Biomolecules that act as catalysts include enzymes, catalytically active antibodies, and ribozymes. An enzyme-catalyzed reaction starts with formation of the enzyme-substrate complex (Chapters 13, 14). From the enzyme-substrate complex, the transition state is reached, leading to the enzyme-product complex that dissociates to enzyme and product (Figure 16.3; see also Scheme 13.9). Enzymes catalyze reactions by binding tightly to the transition state of the catalyzed reaction, leading to its stabilization. We have encountered this phenomenon before as the basis for the efficient inhibition of enzyme activity by transition-state analogs (Section 13.6; Box 13.5). The stabilization of the transition state leads to a decrease in activation energy, both for the forward reaction from reactants to products and for the inverse reaction from products to reactants. The energies of the reactants (enzyme and substrates) and product are unaltered, however, meaning the change in the standard free energy ΔG^0 is identical for the catalyzed and uncatalyzed reaction. Thus, the equilibrium constants of uncatalyzed and enzyme-catalyzed reactions are also identical.

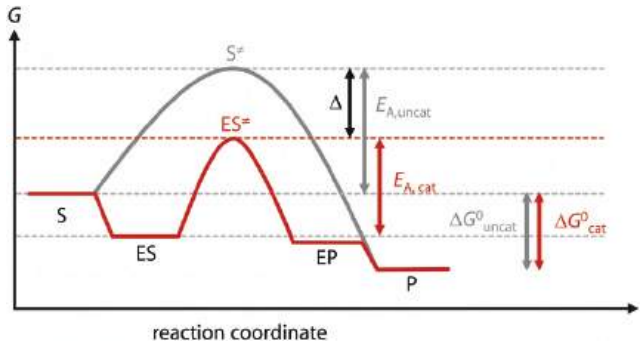


FIGURE 16.3: Enzyme catalysis and transition state stabilization. Enzymes bind their substrate S and form an enzyme-substrate complex (ES). The enzyme-substrate complex is converted to the enzyme-product complex (EP), from which the product P dissociates (see also Scheme 13.9). Enzymes stabilize the transition state (Δ), and achieve an acceleration of the reaction by lowering the energy barrier for both forward (S→P) and reverse reactions (P→S) compared to the uncatalyzed reaction (gray). The energy of the substrate and the product is unaltered, and the standard free energy changes ΔG^0_{cat} and ΔG^0_{uncat} of the catalyzed and uncatalyzed reactions are identical. Hence, enzymes do not alter the equilibrium between substrate and product. Note that the energy diagram reflects the energy levels only for the fraction of substrate that is bound to the enzyme. Under catalytic conditions, the enzyme concentration is much lower than the substrate concentration, and almost all substrate molecules which contribute to the overall energy are in their free form.

Note that for enzyme-catalyzed reactions, we have to distinguish between the equilibrium between substrate and product (K = [P]/[S]), and the *on-enzyme equilibrium* between ES and EP (K_{enz} = [EP]/[ES]). These constants can have different values. Under catalytic conditions, *i.e.* at very low concentrations of enzyme compared to substrate, only a small fraction of the substrate is part of the ES complex; most of the substrate is in the free form. Under these conditions, the on-enzyme equilibrium has very little effect on the overall equilibrium; enzymes present in catalytic concentrations do not alter the overall equilibrium constant of the catalyzed reaction. The equilibrium constant K is defined by the difference in free energy between substrate and product. Under stoichiometric conditions, when the enzyme is present at comparable concentrations to the substrate, the situation is different. Here, most of the substrate is bound to the enzyme, and the on-enzyme equilibrium becomes dominant.

To be able to catalyze a reaction, enzymes need to bind to their substrates to form an ES complex. Many enzymes bind their substrates with high specificity, achieved by forming multiple interactions with the substrate (see Section 18.3.7). High-specificity binding ensures that the enzyme does not catalyze unwanted reactions of similar substrates. High-affinity binding, on the other hand, is not necessarily a useful strategy for catalysis. Extremely tight binding is equivalent to a low energy of the ES complex compared to free substrate. From the energy diagram in Figure 16.3 we can see that the amount to which the energy of the ES complex is reduced directly adds to the activation energy for the enzyme-catalyzed forward reaction, and actually slows down the reaction! To lower the activation energy and accelerate the reaction, enzymes need to bind the transition state of the reaction much more tightly than to the substrate in the ground state (Figure 16.4). Enzymes are often not perfectly complementary to their substrate, but undergo conformational changes that strengthen the interaction with the substrate along the reaction coordinate (induced fit, see Section 18.3.7). The concept of strong binding to transition states as the basis for catalysis has been exploited to generate catalytic antibodies: antibodies generated against transition state analogs of chemical reactions often catalyze the respective reaction, although less efficiently than natural enzymes.

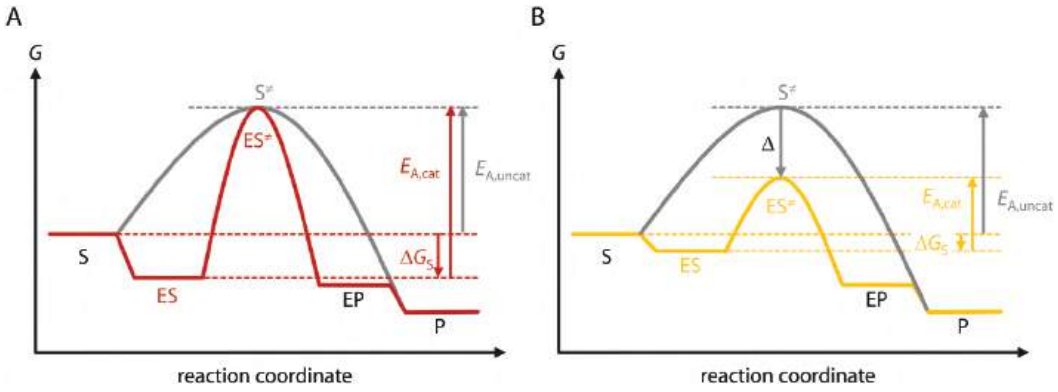


FIGURE 16.4: Effect of substrate and transition state binding on enzyme catalysis. A: An "enzyme" that binds tightly to the substrate (ΔG_S), but not to the transition state does not lower the activation barrier E_a, and does not catalyze the reaction. B: An enzyme that binds weakly to the substrate (ΔG_S), but tightly to the transition state (Δ) lowers the activation barrier E_a for the reaction, and efficiently catalyzes the conversion from substrate to product.

Very tight binding of the enzyme to the products is also counterproductive for efficient catalysis because the products need to dissociate for regeneration of the enzyme. Tight binding of the substrate and a concomitant low energy of the EP complex would render substrate dissociation endergonic, and would prevent the enzyme from multiple turnovers. This problem is often encountered in catalytic RNAs, such as ribozymes that cleave an RNA substrate. Here, the cleaved RNA strand remains bound to the ribozyme by base pairing and does not readily dissociate. The dissociation

then becomes rate-limiting for the overall reaction. The catalytic efficiency of protein enzymes is also reduced when substrate and product are very similar. In this case, the product is also bound tightly, leading to product inhibition (Section 13.6.1). An example is ATP hydrolysis, where the substrate ATP and the product ADP/P_i are chemically not too different. Many ATP-hydrolyzing enzymes, such as the muscle protein myosin, show strong product inhibition by ADP.

In the following, we will briefly discuss the main strategies of enzymes to stabilize transition states and to accelerate biochemical reactions. For more detailed information, the reader is referred to more specialized textbooks on catalysis.

16.2.1 Acid-base catalysis

Acid-base catalysis involves the transfer of a proton to the reactant to stabilize negative charges in the transition state (acid catalysis), or the abstraction of a proton to stabilize a developing positive charge (base catalysis). The overall reaction scheme for catalysis of the reaction of A and B to product P by an acid HX is (Scheme 16.1):

$$\mathrm{A + H\text{-}X \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AH^+ + X^-}$$
$$\mathrm{AH^+ + B \xrightarrow{k_2} P}$$

scheme 16.1

The overall reaction velocity of product formation is

$$v = k_2\left[AH^+\right][B] \qquad \text{eq. 16.14}$$

Similar to the steady-state approximation for the ES complex in the derivation of the rate equation for enzyme reactions (Section 13.2), we can formulate a steady-state condition for $[AH^+]$:

$$\frac{\mathrm{d}\left[AH^+\right]}{\mathrm{d}t} = k_1[A][HX] - k_{-1}\left[AH^+\right]\left[X^-\right] - k_2\left[AH^+\right][B] \approx 0 \qquad \text{eq. 16.15}$$

From eq. 16.15, we can derive an expression for $[AH^+]$

$$\left[AH^+\right] = \frac{k_1[A][HX]}{k_{-1}\left[X^-\right] + k_2[B]} \qquad \text{eq. 16.16}$$

and substitute $[AH^+]$ in the reaction velocity (eq. 16.14) to obtain eq. 16.17:

$$v = \frac{k_1 k_2[A][HX][B]}{k_{-1}\left[X^-\right] + k_2[B]} \qquad \text{eq. 16.17}$$

In a first case, we assume that protonation and deprotonation are rapid, and product formation is rate-limiting, with

$$k_{-1}\left[X^-\right] \gg k_2[B] \qquad \text{eq. 16.18}$$

so that we can neglect the term $k_2[B]$ in the denominator in eq. 16.17, and can approximate the reaction velocity v as

$$v \approx \frac{k_1 k_2[A][HX][B]}{k_{-1}\left[X^-\right]} \qquad \text{eq. 16.19}$$

From the definition of the equilibrium constant K_A for the protolysis of HX

$$K_A = \frac{[H^+][X^-]}{[HX]} \qquad \text{eq. 16.20}$$

we obtain

$$\frac{[HX]}{[X^-]} = \frac{[H^+]}{K_A} \qquad \text{eq. 16.21}$$

which we can now substitute into eq. 16.19, which gives

$$v \approx \frac{k_1 k_2 [A][B][H^+]}{k_{-1} K_A} \qquad \text{eq. 16.22}$$

The reaction velocity thus depends linearly on the concentration of protons. *Specific acid catalysis* refers to such reactions where protons act as the catalysts. In analogy, *specific base catalysis* refers to reactions catalyzed by hydroxide ions. Examples are the base-catalyzed cleavage of ribonucleic acid, or the acid-catalyzed depurination of DNA (Box 16.1).

BOX 16.1: SPECIFIC ACID-BASE CATALYSIS.

The cleavage of RNA at high pH occurs *via* specific base catalysis by hydroxide ions. DNA undergoes depurination under acidic conditions *via* specific acid catalysis by protons (Figure 16.5).

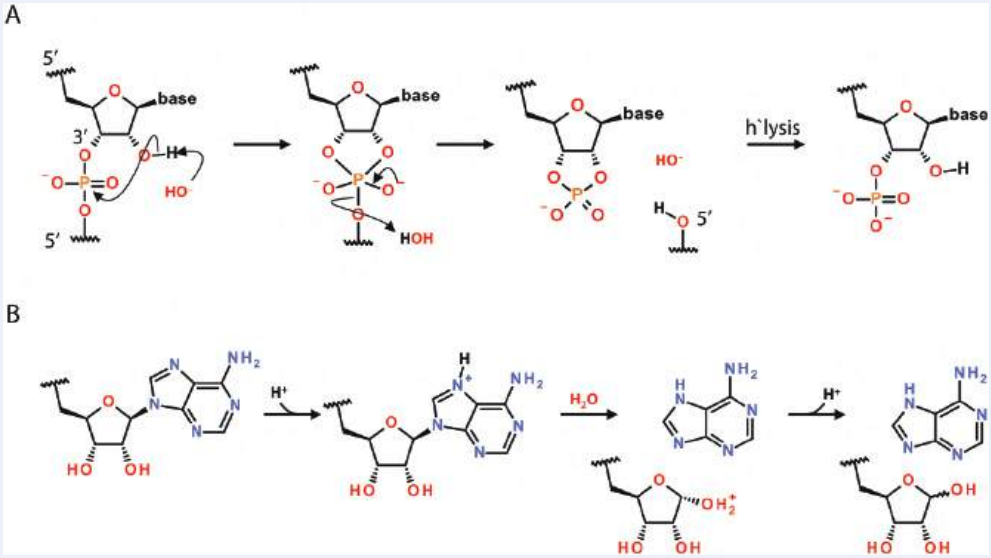


FIGURE 16.5: Specific acid and base catalysis. A: RNA cleavage by specific base catalysis (OH^-). B: DNA depurination by specific acid catalysis (H^+).

Now we treat the case

$$k_{-1}\left[X^{-}\right] \ll k_2[B] \qquad \text{eq. 16.23}$$

meaning a slow protolysis equilibrium is followed by rapid product formation. In this case we can neglect the first term in the denominator of eq. 16.17, and obtain a different approximation for v:

$$v \approx \frac{k_1 k_2[A][HX][B]}{k_2[B]} \qquad \text{eq. 16.24}$$

or

$$v \approx k_1[A][HX] \qquad \text{eq. 16.25}$$

Here, the rate of the reaction is proportional to the concentration of the acid, not to the proton concentration. This case is called *general acid catalysis*. Specific and general acid catalysis are associated with different energy profiles (Figure 16.6).

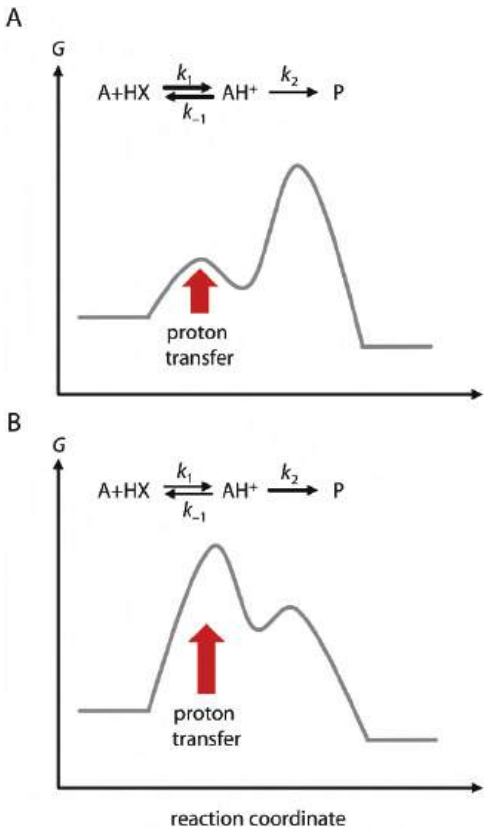


FIGURE 16.6: Energy diagrams for general and specific acid catalysis. A: In specific acid (base) catalysis, rapid transfer of the proton is followed by a slow conversion of the protonated substrate to product(s). The rate constant of the overall reaction depends linearly on the proton concentration, and is thus pH-dependent. B: In general acid (base) catalysis, proton transfer is slow, and dissociation of the protonated substrate to products is rapid. The rate constant of the overall reaction is proportional to the concentration of the free acid HA.

In general acid-base catalysis, positive or negative charges developing during the reaction, *i.e.* in a transition state, are stabilized by proton transfer to a base or from an acid, respectively. This is a common mechanism in enzyme-catalyzed reactions (Box 16.2).

BOX 16.2: GENERAL ACID-BASE CATALYSIS.

The enzyme RNase A catalyzes the hydrolysis of phosphodiester bonds in ribonucleic acids (Figure 16.7). In the first step, a histidine side chain (His12) acts as a general base, and abstracts a proton from the 2'-hydroxyl of the scissile bond. The deprotonated 2'-O^- then forms a cyclic phosphate by nucleophilic substitution at the phosphorous. The leaving group, the 5'-end of the cleaved RNA, is protonated at the 5'-O^- by a second histidine, His 119, which acts as a general acid. In the second step, the inverse reaction takes place with water: His119 now acts as a general base, and accepts a proton from H_2O. The hydroxyl then performs a nucleophilic substitution at the phosphorus of the cyclic phosphate. The 2'-O^- leaving group is protonated by His12 that now acts as a general acid. Catalytic residues for acid-base catalysis include histidine, serine, and cysteine, but also cofactors, e.g. pyridoxal phosphate.

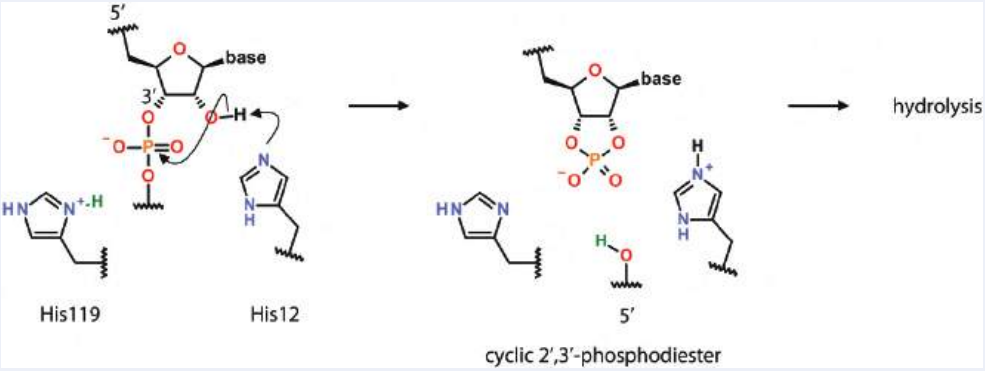


FIGURE 16.7: General acid-base catalysis in the reaction of RNase A. His12 acts as a general base and abstracts a proton from the 2'-OH of the ribose. The 2'-O^- then performs a nucleophilic substitution at the phosphorous. His119 acts as a general acid and protonates the leaving group. The 5'-RNA dissociates from the enzyme, and is replaced by a water molecule. The hydrolysis of the cyclic phosphodiester is achieved by the reverse sequence: His119 acts as a general base and accepts a proton from the water molecule. The OH^- performs a nucleophilic substitution at the phosphorous, and His12 acts as a general base and protonates the leaving group, the 2'-O^- of the ribose. At the end of the reaction, His12 and His119 are in the same protonation states as before.

16.2.2 Electrostatic and covalent catalysis

Charges of the transition state can also be stabilized by electrostatic interactions (see Section 17.2.2): positive charges can be stabilized by negatively charged groups in the vicinity (*nucleophilic catalysis*), and negative charges by positively charged groups or metal ions (*electrophilic catalysis*). A common mechanism for the enzymatic catalysis of biochemical reactions involves *covalent catalysis* (Box 16.3), in which the catalyst, often a prosthetic group or a reactive protein side chain, activates the substrate by formation of a covalent intermediate. Active site residues for covalent catalysis are nucleophilic, and are good leaving groups to ensure turnover of the intermediate to products. Pyridoxal phosphate-dependent enzymes such as aminotransferases are prominent examples of covalent catalysis (Box 16.3). Serine proteases such as trypsin and chymotrypsin also employ covalent catalysis.

BOX 16.3: ACTIVATION BY FORMATION OF A COVALENT INTERMEDIATE.

Aminotransferases are pyridoxal phosphate-dependent enzymes that transfer amino groups from amino acids to α-keto acids during amino acid degradation (Figure 16.8). Aspartate aminotransferase catalyzes the transfer of the amino group from aspartate to α-ketoglutarate, leading to formation of oxaloacetate and glutamate. In a reaction catalyzed by glutamate dehydrogenase, glutamate then undergoes oxidative deamination, which leads to regeneration of α-ketoglutarate and the formation of ammonium ions. In mammals, ammonium ions are converted to urea that is excreted. Aminotransferase contains the prosthetic group pyridoxal phosphate that forms a covalent bond with a lysine residue in the active site (Figure 16.8). When the substrate amino acid binds, a covalent intermediate between the amino acid and pyridoxal phosphate is formed, in which the amino acid replaces the lysine residue. The resulting aldimine intermediate is isomerized to a ketimine intermediate that is hydrolyzed to a α-keto acid and pyridoxamine phosphate. The α-keto acid dissociates from the enzyme. In the second half of the reaction, the binding site is occupied by α-ketoglutarate that forms a ketimine with pyridoxamine phosphate. Now the ketimine is isomerized to an aldimine that hydrolyzes to glutamate and pyridoxal phosphate.

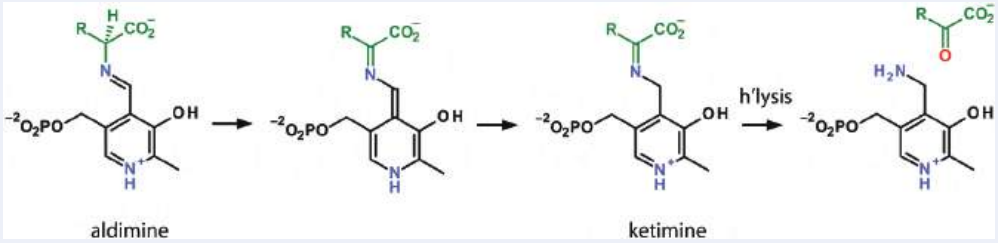


FIGURE 16.8: Covalent catalysis by aminotransferases. Aminotransferases use pyridoxal phosphate as a prosthetic group. Pyridoxal phosphate forms a covalent bond with a lysine residue in the active site of the aminotransferase. An amino acid that is bound as a substrate replaces the catalytic lysine, and forms an aldimine with pyridoxal phosphate that is isomerized to a ketimine, and hydrolyzed to the α-keto acid and pyridoxamine phosphate.

16.2.3 Intramolecular catalysis and effective concentrations

Intramolecular catalysis has a major advantage in that it reduces the entropic cost of bringing together reactants. This effect relates to the entropic term in the transition state theory (eq. 15.12). When two substrate molecules react to one product molecule, an overall reduction in entropy is observed. This entropy loss is only compensated slightly by increases in entropy caused by different modes of internal rotation and vibration in the product molecule. By using intramolecular catalysis, enzymes avoid this unfavorable entropic term. A reaction with general acid- or base-catalysis has a second-order rate constant, and its rate increases linearly with the concentration of the acid or base (eq. 16.25). Catalysis of the same reaction within an enzyme-substrate complex is a unimolecular reaction, with a first-order rate constant. As the catalytic group and the substrate are part of the same entity, the concentration of the acid or base is difficult to define. An "effective" concentration can be calculated by comparing reaction rates of model compounds that contain the substrate and the catalytic group within the same molecule with rates of the intermolecular reaction at different concentrations of the acid-base. The concentration of external acid or base required to give the same reaction rate as observed for intramolecular catalysis corresponds to the effective concentration of this group in the enzyme. Effective concentrations are typically in the high molar range (general acid catalysis) and can reach values as high as 10^5 M in nucleophilic catalysis. These high local concentrations contribute to the high catalytic efficiency of enzymes.

QUESTIONS

16.1 Derive the Langmuir-Hinshelwood binding isotherm and the relationship for the reaction velocity v of a heterogeneously catalyzed, bimolecular reaction, *i.e.* for the conversion of two substrate molecules S to a product P.

16.2 Why is histidine often present in active sites of enzymes?

16.3 Using a thermodynamic cycle, illustrate why an enzyme needs to bind tightly to the transition state, not the ground state of the reaction, to be an efficient catalyst.

16.4 An enzyme reduces the activation energy E_a of the catalyzed reaction A ⇄ B by 5.7 kJ mol^{-1}. What is the acceleration of the forward and reverse reaction at 25°C?

16.5 An enzyme-catalyzed reaction proceeds from substrate to product though a transition state, with an activation energy E_a = 12 kJ mol^{-1}. The reverse reaction is two-fold faster than the forward reaction. Calculate the activation energy for the reverse reaction, the rate constants for forward and reverse reaction, the equilibrium constant and the free energy difference between substrate and product (pre-exponential factor = 1, T = 298 K).

REFERENCES

De La Cruz, E. M., Sweeney, H. L. and Ostap, E. M. (2000) ADP inhibition of myosin V ATPase activity. *Biophys J* **79**(3): 1524–1529.
· product inhibition of myosin

Jencks, W. P. (1987) *Catalysis in Chemistry and Enzymology*. New York, Dover.
· a compendium about the principles of catalysis.

Knowles, J. R. (1991) Enzyme catalysis: Not different, just better. *Nature* **350**(6314): 121–124.
· a review article that explains why enzymes are excellent catalysis

Moore, S. A. and Jencks, W. P. (1982) Reactions of acyl phosphates with carboxylate and thiol anions. Model reactions for CoA transferase involving anhydride formation. *J Biol Chem* **257**(18): 10874–10881.
· effective concentrations

Page, M. I. and Jencks, W. P. (1971) Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc Natl Acad Sci U S A* **68**(8): 1678–1683.
· entropic effects in catalysis

Pollack, S. J., Jacobs, J. W. and Schultz, P. G. (1986) Selective chemical catalysis by an antibody. *Science* **234**(4783): 1570–1573.
· description of the first catalytic antibody that tightly binds to a transition state analog



PART III

Molecular structure and stability

Parts I and II have introduced the energetic and kinetic principles that govern molecular interactions. These principles are the key for understanding molecular structure and stability and molecular recognition. While structure and stability of biopolymers are based on intramolecular interactions, intermolecular interactions are the key for ligand binding and molecular recognition in general.

Molecular recognition is at the heart of biology, *e.g.* enzymology and catalysis (Chapters 13–16), ligand binding (Sections 18.3.7, 19.5.4.2, and ligand-detection techniques in Part IV), protein-protein- (Section 18.3.6) and protein-nucleic acid interactions (Section 19.5.4). All of these interactions follow the rules of thermodynamics: they seek to reach the state of the system with the lowest free energy.

17

Molecular structure and interactions

In this Chapter, we discuss the structural aspects and biophysical consequences of molecular interactions, starting from simple organic entities of a few atoms to increasingly large biomolecules with masses in the MDa range. Although some basic concepts of organic chemistry are reviewed, the reader should be familiar with partial charges and electronegativity, hybridization of atomic orbitals to molecular orbitals (sp, sp^2, sp^3) and how these dictate stereochemistry, the different bond types (σ, π), and bond energies.

17.1 ISOMERS: CONFIGURATION AND CONFORMATION

The biophysical properties of molecules are determined by their atomic composition, the connectivity of these atoms, and the resulting shape. The shape of molecules is of high biological importance when it comes to molecular recognition and catalysis. Molecules with the same atomic composition can exist in different forms. The collective term for all forms of a molecule with the same sum formula is *isomers*. Isomers play an important role in biology, which is why we briefly recapitulate here the different kinds of isomers that molecules can form (Figure 17.1).

Conceptually, isomers can be divided into two major groups, those with the same connectivity of atoms by bonds, termed *stereoisomers*, and those with fundamentally different connectivity, termed *constitutional isomers*. A special form of constitutional isomers are the *tautomers* that only differ in the position of a single hydrogen atom. Constitutional isomers can only be converted from one to another by bond breakage. This is not necessarily true for stereoisomers. Here, we need to distinguish between *conformations* that can be converted into one another without breaking bonds, and *configurations* whose interconversion requires bond breakage. Subgroups of configurations are the *enantiomers*, pairs of molecules related to each other by mirror symmetry, and *diastereomers*, molecules with more than one stereocenter but no mirror symmetry. Diastereomers comprise *epimers*, enantiomers that only differ in the configuration at one stereocenter, and *anomers*, a special case of diastereomers in sugars with different configurations at C1.

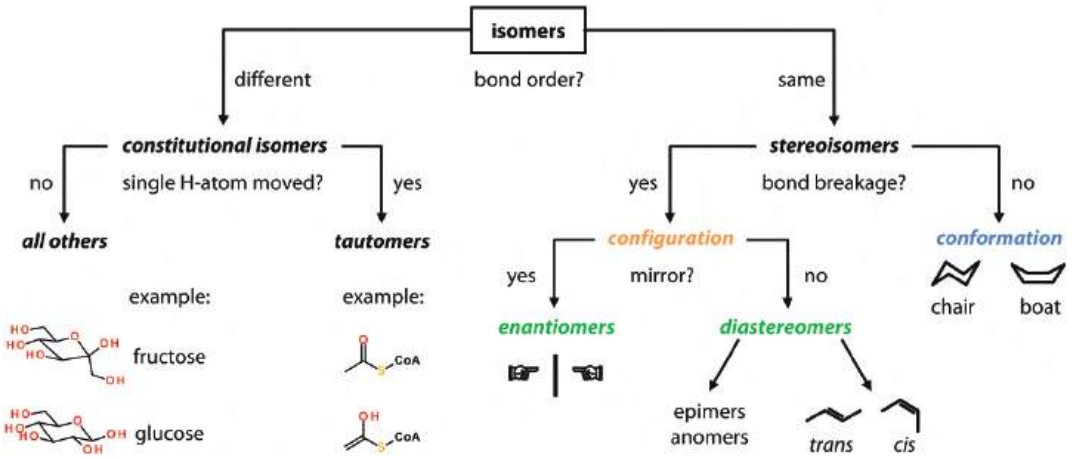


FIGURE 17.1: Classification of isomers. Isomers have the same sum formula. Molecules with a different bond order are constitutional isomers, all others are stereoisomers. Two constitutional isomers that just differ in the position of a single hydrogen atom are tautomers. The stereoisomers on the right can be grouped into isomers with different conformation (no bond breakage necessary for interconversion) or configuration (bond breakage necessary). Conformers interconvert by bond rotations and twisting movements. Enantiomers are mirror images of one another, whereas diastereomers include both *cis/trans* isomers and isomers with more than one stereocenter.

The interconversion of constitutional isomers and of different configurations requires bond breakage, which is associated with large changes in free energies in the hundreds of kJ mol^{-1} (Table 17.1). In the cell, the interconversion of these isomers is typically catalyzed by enzymes with breakage and re-formation of chemical bonds and rearrangements at physiological conditions.

Isomers have a different three-dimensional arrangement of their constituting atoms, and thus different distributions in space of groups of different size, polar and non-polar groups, and hydrogen bond donors and acceptors. As a consequence, they differ substantially in their chemical and biophysical properties. In the following, we will discuss a few illustrative examples of isomers that are prominent in biology. We will start with configurations (Section 17.1.1), and then turn to conformations (Section 17.1.2).

17.1.1 Configurations

Pairs of constitutional isomers are frequently found in metabolism. Examples are the sugars glucose and fructose or the two important glycolysis metabolites dihydroxyacetone phosphate and glycerol aldehyde-3-phosphate (Figure 17.1). The enzymes that interconvert these pairs are glucose isomerase and triosephosphate isomerase, respectively, (Section 2.1.2). The molecules in each pair have very different structures and are recognized by different metabolic enzymes in very different pathways.

Tautomers, constitutional isomers that only differ in the position of a single hydrogen atom, are ubiquitous in biology. The energy difference between two tautomers is low compared to isomers that require other bonds to be broken, and the different tautomers may co-exist at appreciable concentrations in equilibrium. If the proton shifts from a carbon atom next to a keto, aldehyde, or ester group to the oxygen atom of that group (Figure 17.2), we speak of *keto/enol tautomerism*. It is part of interconversion of metabolites in many metabolic pathways, and frequently observed in reducing sugars (Figure 17.1). Keto/enol tautomerism can have tremendous effects on the reactivity of the molecule. For example, acetyl-CoA, an activated thioester, is an electrophile in its keto form but a nucleophile in its enol form. If hydrogen atoms move from a carbon atom to a neighboring nitrogen atom, the isomers are called *imino/enamine tautomers*. The side chain of histidine (Figure 17.2) often undergoes imino/enamine tautomerism. The hydrogen bonding capabilities of the two isomers are

different, enabling different interactions. Similarly, the imino and enamine forms of nucleobases have different hydrogen bonding capabilities (Section 19.1).

Configurational isomers that can only be interconverted by bond breakage are also ubiquitous in biology. Enantiomers are probably the most distinctive configurational isomers in biological molecules (Figure 17.2). These molecules possess a center of mirror symmetry or handedness, which is somewhat confusingly also called *stereogenic center*, *stereocenter*, or *chiral center*. Entire classes of molecules, such as amino acids in proteins, nucleotides in nucleic acids, and lipids in membranes, occur only in one enantiomeric form in nature. Most chiral centers in biological molecules are carbon atoms with four different substituents R_1–R_4. To establish whether a molecule has the *(R)* (lat.: *recte*, right) or (*S*) (lat.: *sinister*, left) configuration, the substituents are prioritized by their atomic numbers, multiplied by the bond order, if applicable. The molecule is turned such that the substituent R_1 of lowest priority points to the background. Going from R_2 to R_4 in a clockwise (right) or counter-clockwise (left) rotation then defines *(R)* and *(S)* configurations.

Molecules that contain more than one stereocenter do not exhibit mirror symmetry. This is the case for all carbohydrates, nucleotides, most lipids, and the amino acids isoleucine and threonine. The presence of *n* stereocenters gives rise to 2^n diastereomers in 2^{n-1} enantiomeric pairs. In biology usually only a single diastereomer is relevant, *i.e.* a single combination of (R) and (S) stereocenters. Molecules that contain several stereocenters and that differ only in the configuration of a single chiral center are termed *epimers*. For example, mannose is the C2-epimer of glucose, and galactose and glucose are C4-epimers. *Anomers* are a special case of epimers that are generated during ring formation of sugars. For example, glucose exists as α- and β-glucose, which are epimers at C1. In solution, anomers interconvert spontaneously.

Cis/trans isomers, a subgroup of diastereomers, can be interconverted by breaking a π-bond and re-establishing it after a 180° rotation about the remaining σ-bond (Figure 17.2). The *trans* isomer is usually more stable than the *cis* isomer because the chemical groups joined by the double bond are further apart, minimizing steric hindrance. *Cis/trans* isomerization is often associated with a large change in shape. For example, the *cis*-to-*trans* isomerization of retinal, a key step in the visual process (Box 21.1 in Section 21.2.6), changes the shape of this molecule so much that it dissociates from its binding partner. Another example is unsaturated fatty acids, such as oleic acid, where *cis* or *trans* isomers have different shapes. As a consequence, the *cis* isomers cannot pack as tightly as the corresponding *trans*- and saturated fatty acids, and a higher content of *cis*-fatty acids can therefore increase the fluidity of membranes (Section 18.4).

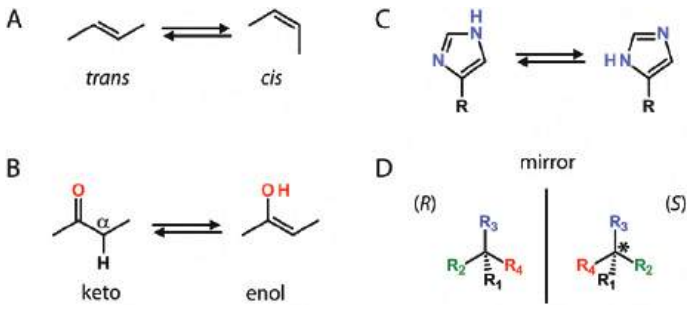


FIGURE 17.2: Common configurational isomers. A: In *cis/trans* isomerism, a π-bond is broken and re-formed after rotation about the remaining σ-bond. B: In keto-enol tautomers, a hydrogen atom is shifted between a carbon atom and an oxygen atom. C: A substituted imidazole, such as histidine, can also exist as two tautomers. Both are observed in proteins, but the one shown on the left is more stable by about 14 $kJ \cdot mol^{-1}$. D: Enantiomers possess mirror symmetry (shown as a vertical line). The most common stereocenters in biology are asymmetric carbon atoms with four different substituents R_1-R_4.

17.1.2 Conformations

We now turn from configurations, which are different chemical entities, to *conformations*, which are different three-dimensional arrangements of atoms in the same molecule. Similar to different configurations, conformations of a molecule may have very different three-dimensional shapes. The shape of individual conformers is established by individual arrangement of groups of different size, polar and non-polar groups, and hydrogen bond donors and acceptors. Conformer shapes dictate affinity and specificity in molecular recognition, thereby playing a profound role in biology. Different conformations in biology range from small-scale to large-scale, from rotamers of amino acid side chains or nucleotide conformations to protein (Section 18.3) and nucleic acid conformations (Sections 19.5 and 19.6). In essence, conformations determine the stability of biomolecules.

The set of all conformations a molecule can adopt is known as *conformational isomers* or *conformers* for short. Changes in conformation are brought about by rotations about single bonds and transient distortions of bond angles, which is usually associated with relatively low energy barriers of < 50 kJ·mol^{-1}. As a consequence, the interconversion of conformers is a rapid process, and at ambient temperate a molecule can sample thousands of conformations per second (Table 18.12).

17.2 The configurational and conformational energy of molecules

Molecular structures are fully described by lists of atom coordinates (x,y,z). Different isomers have different positions of the constituting atoms, and differ in their free energy. The total internal energy E_{tot} of an isolated molecule depends on all intramolecular interactions present, which can be divided into covalent and non-covalent contributions:

$$E_{tot} = E_{covalent} + E_{non\text{-}covalent} \qquad \text{eq. 17.1}$$

Covalent or *bonded* interactions are defined between up to four atoms connected by three chemical bonds (Section 17.2.1). Covalent interactions are different for different configurations, but identical (or at least very similar) for different conformations. Non-covalent or *non-bonded* interactions occur between atoms more than three bonds away in sequence. They include all other types of interactions within the molecule, including electrostatic attraction or repulsion between ions and dipoles, and steric repulsions (Section 17.2.2). Non-covalent interactions are different for different configurations and conformations. In the following, we will show how we can describe the energies of different conformers of a molecule by summing the energies of covalent and non-covalent interactions according to eq. 17.1. The same principles apply to the description of different configurations of molecules, and can be extended to interactions between molecules.

17.2.1 Covalent interactions

Covalent bonds define the spatial arrangement and connectivity of atoms in a molecule. To formulate its total energy, we can describe covalent or *bonded* interactions by energy terms that depend on the positions of bonded atoms. These energy terms include contributions from the atomic bonds, the bond angles, and the torsion angles (eq. 17.2).

$$E_{covalent} = E_{bond} + E_{angle} + E_{torsion} \qquad \text{eq. 17.2}$$

Atomic bonds include single, double, and triple bonds. In general, the shorter the bond and the higher the bond order and the more polar a bond is, the more stable it is (Table 17.1). Bond angles or *1,3-angles* are defined by three atoms connected by two bonds. *Torsion angles*, also termed *1,4-* or *dihedral angles*, are defined by four atoms that are linearly connected by three bonds. Figure 17.3 shows the shapes of the energy functions for changes in bond length, 1,3-angles, and torsion angles.

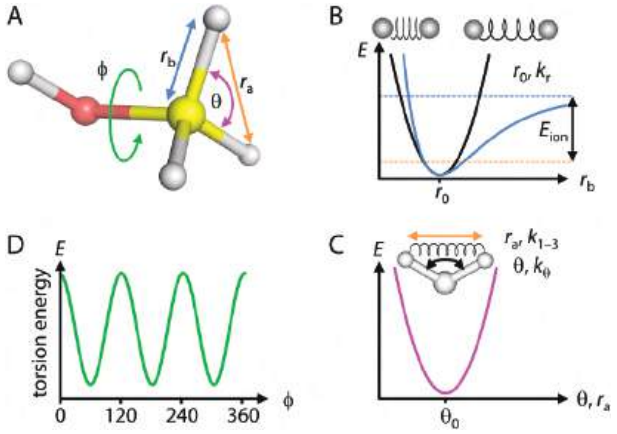


FIGURE 17.3: Energy terms describing covalent bonds and their dependence on distance and angles. A: The bond distance r_b, the 1,3-angle θ with its distance r_a, and the 1,4-angle (torsion) ϕ are indicated in separate colors for the simple molecule methanol (CH_3OH). The plots in B-D show the energy variation with distance or angle in the same color code. B: For the equilibrium distance r_0, a compression shows a parabolic effect on energy (black curve), but this potential function is inaccurate away from r_0, particularly for large values of *r*. Lengthening of bonds will eventually lead to bond breakage. The Morse potential (blue) has an asymptotic trace for $r \gg r_0$ (blue dashes). k_r is the spring constant. The ground state with the zero-point energy is indicated by the orange dashes; the difference between the ground state and the asymptote is the ionization energy E_{ion}. C: The energy change upon compression or widening of the equilibrium bond angle θ_0 can be described by a parabola (magenta). Measuring the distance between two atoms that are bonded to a third atom is an equivalent way of describing the bond angle (orange). k_θ and k_{1-3} are the spring constants. D: The torsion angle energy of a methyl group follows a sinusoid pattern with minima repeating every 120° (green).

17.2.1.1 Covalent bonds

The energy of a covalent bond depends on the elements involved and on the bond order (Table 17.1). Probably the simplest model to describe the energy potential of a covalent bond is the harmonic oscillator: a spring with a spring constant k_r acting between the atoms. This concept is treated in detail in Section 21.4.1 on infrared spectroscopy. For our discussion here, it is sufficient to note that integration of Hooke's law (eq. 21.51, eq. 21.53) gives the parabolic potential energy as:

$$E_{\text{bond}} = \frac{1}{2} k_r \left(r - r_0\right)^2 + \textit{offset} \qquad \text{eq. 17.3}$$

The minimum bond energy corresponds to the equilibrium bond distance r_0. The "offset" in eq. 17.3 accounts for the zero-point energy in the ground state, the energy difference between a quantum-mechanical system and a system described with classical mechanics. The energy is symmetric about r_0, and any positive or negative deviation from r_0 increases the energy by the same amount; eq. 17.3 is a *harmonic* function. The harmonic potential describes the bond energy well for small deviations from r_0 in both directions, and is a good approximation for $r < r_0$. The description fails for $r \gg r_0$, however: the harmonic model predicts that the energy would increase to infinity upon excessive lengthening of the bond, which is not true because at some point the bond will break and the energy of the system will then remain constant. Hence, for larger bond distances the potential function has to be modified to take into account the finite energy of the system once the bond is broken (Figure 17.3). This shape is reflected in the *Morse potential*:

$$E_{\text{bond}} = D_e \cdot \left(1 - e^{-\sqrt{\frac{k_r}{2D_e}} \cdot (r - r_0)}\right)^2 + \textit{offset} \qquad \text{eq. 17.4}$$

D_e is the depth of the energy well and depends on the nature of the bonded atoms. For large distances r, E_{bond} approaches D_e. The dissociation energy is a little less than D_e because in the ground state of a molecule the energy of the bond is not zero. The energy range of covalent bonds is 200–800 kJ mol^{-1}, much higher than any non-covalent interactions, which amount to a few kJ mol^{-1}. Table 17.1 lists mean energies and lengths of frequently encountered chemical bonds in biomolecules.

TABLE 17.1:
Mean energies and lengths for bonds found in biological molecules.

bond	energy (kJ mol^{-1})	bond length (pm)
C–H	415	109
O–H	465	96
N–H	389	101
S–H	348	143
C–C	348	154
C=C	611	147
C≡C	837	138
C–O	360	143
C=O	~740*	121
C–N	306	147
C=N	615	128
C≡N	892	114
C–S	272	182
S–S	251	203
S=O	578	144/150**

* Aldehydes: 737 kJ mol^{-1}. Ketones: 749 kJ mol^{-1}. **Sulfones and sulfoxides. Data from *Handbook of Chemistry and Physics* (hbcpnetbase.com), with permission.

17.2.1.2 Bond angles and torsion angles

Similar to the distance dependence of the energy of single covalent bonds, the energy of two bonds is determined by their bond angle. Small changes in the bond angle around its equilibrium value can also be described by a harmonic potential with the spring constant k_θ and the equilibrium bond angle θ_0:

$$E_{angle} = \frac{1}{2} k_\theta (\theta - \theta_0)^2 = \frac{1}{2} k_{1-3} (r - r_0)^2 \qquad \text{eq. 17.5}$$

Alternatively, bond-angle variations can be described by the oscillation of a spring with the spring constant k_{1-3} that connects the first and third atom of the three atoms defining the bond angle θ and the distance r between these two atoms (Figure 17.3). The expression of bond angles in terms of distances is equivalent to angles, but distances are more convenient parameters for computation (Sections 31.9 and 31.13).

The angle enclosed by three bonded atoms 1–2–3 can be calculated from the atom coordinates. Specifically, the magnitudes and the dot product of two vectors are required, namely the vector **u** connecting atoms 1 and 2, and the vector **v** connecting atoms 2 and 3. The components of **u** and **v** are simply the differences of the (x,y,z) coordinates of the atom pairs between which the vector is drawn. The angle θ between the two vectors is then calculated according to eq. 17.6

$$\theta = \arccos \frac{u_1 v_1 + u_2 v_2 + u_3 v_3}{|\mathbf{u}||\mathbf{v}|} \qquad \text{eq. 17.6}$$

where $|\mathbf{u}|$ and $|\mathbf{v}|$ are the magnitudes of the vectors calculated using the Pythagoras formula (eq. 31.84) and the numerator is the dot product.

Bond angles define the angle between the two bonds between three bonded atoms 1–2–3. Torsion angles, also known as dihedral angles or 1,4-angles, describe the spatial relation of four

consecutively bonded atoms 1–2–3–4. Rotation about the central bond (2–3) changes the relative positions of atoms 1 and 4. If viewed along the central bond, the second and third atoms are projected onto each other, leaving three "atoms" or reference points that describe the torsion angle in the same way as the 1,3-angle (Figure 17.4). In principle, the values for torsion angles could be calculated using eq. 17.6 in a similar way as for bond angles by determining the angle between the vector **u** connecting atoms 1 and 2 and vector **v** between atoms 3 and 4. However, torsion angles have a direction: they can be positive (clockwise) or negative (anti-clockwise). By definition, the torsion angle is the smaller value when looking along the central 2–3 bond and rotating atom 1 towards atom 4 (Figure 17.4). Their values range between –180° and +180°. To obtain the sign, or handedness, from the calculation, the torsion angle is defined as the angle between the vectors normal to the planes spanned by atoms 1–2–3 and 2–3–4. This calculation involves the cross product (Section 31.9.2). To visualize torsions, two different projections are in use, the Newman and the sawhorse projection. The Newman projection emphasizes the direction of the torsion by projecting the molecule along the 2–3 bond. The sawhorse projection emphasizes the dihedral angle by showing all four atoms such that one can imagine the two planes 1–2–3 and 2–3–4 (Figure 17.4).

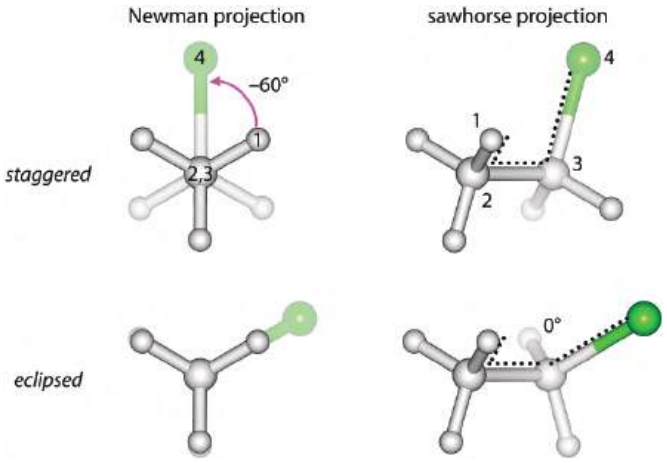


FIGURE 17.4: Staggered and eclipsed conformation of chloroethane. The top panel shows the staggered conformation with all substituents maximally apart. The torsion angle is –60°, regardless of which direction of view (2–3 or 3–2) we take. On the bottom, the eclipsed conformation is drawn with a torsion angle of 0°. The left-hand side is the Newman projection, the right-hand side is the sawhorse projection. Both projections show the same conformations with different emphasis. The Newman projection reduces the 1,4-angle to the 1,3-angle concept. The sawhorse projection emphasizes the bond about which the substituents are rotated.

Depending on the arrangement of substituents, conformations are called either *staggered* or *eclipsed* (Figure 17.4). In the staggered conformation, the substituents are positioned with the maximum possible distance from each other, which corresponds to an energetic minimum. In contrast, the eclipsed conformation brings the substituents closer together, and steric clashes give rise to a maximum of energy. Staggered and eclipsed conformations are the energetic extremes, and the many conformations between these extremes have intermediate energies. *Gauche* are staggered conformations with torsional angles of 60°, where the largest substituents are further away from other atoms than in any *syn*-conformation. In the staggered conformation with a torsional angle of 180°, termed *anti*, the largest substituents are furthest apart, corresponding to the conformation with the lowest energy. As a general rule, all conformations in biological molecules attempt to be staggered, and ideally they are *anti*. Torsion energies larger than the minimum are known as the *torsional strain* of a molecule. Because torsion angles are so variable, they play an important role in defining the preferred conformations of flexible molecules such as fatty acids (Section 18.4), but also of protein and nucleic acid backbones. The torsional energy of individual four-atom configurations may differ by only a few kJ mol^{-1}, but in large macromolecules, these energies will sum to large values.

The energy as a function of the torsion angle reflects the proximity between the substituents 1 and 4, which is periodic. The torsional energy can be described by a cosine function (eq. 17.7 and Figure 17.5).

$$E_{\text{torsion}} = \frac{1}{2} k_{\phi} \left(1 + \cos\left(n\phi - \delta\right)\right) \qquad \text{eq. 17.7}$$

The periodicity n indicates the number of energy minima within a 360° rotation about the bond. The variable δ is the phase shift of the cosine function, and the spring constant is k_{ϕ}. For two sp^3-hybridized tetrahedral carbon atoms, n equals three and δ is zero (Figure 17.5). The energy minima are therefore at ϕ angles of 60°, 180°, and 300° (–60°). The configuration of four atoms 1–2–3–4 with atoms 2 and 3 connected by a double bond (n = 2 and δ = 180°) may be treated as a special case of a torsion angle with distinct energy minima at 180° (*trans*) and 0° (*cis*). A *cis/trans* isomerization of a double bond requires breaking of the π-bond by rotating around the σ-bond, which leads to an energy maximum 90° away from planarity (Figure 17.5).

For molecules with the same substituents 1 and 4 connected to the central atoms 2–3, all energetic minima and maxima are equal. In molecules with different substituents 1 and 4, this is not the case. Here, upon rotation of the central 2–3 bond, different distances between atoms 1 and 4 are sampled and the torsion potential function will become more complicated (Figure 17.5). This is the common situation in biological molecules.

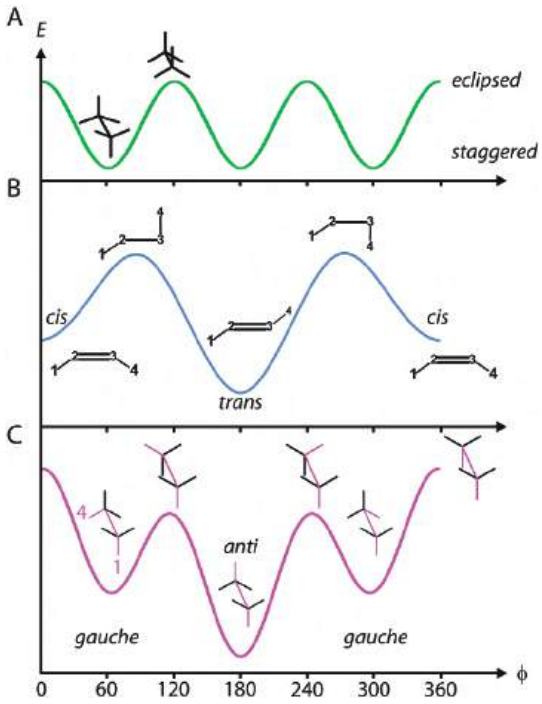


FIGURE 17.5: Periodicity of the conformational energy upon rotation about a dihedral angle. A: This panel shows a single bond with all substituents the same, as in ethane. Every 60°, the torsion energy oscillates between the low-energy, staggered conformation and the high-energy, eclipsed conformation. The energy difference between these conformations is about 12 kJ mol^{-1}. B: Rotation about a double bond has a periodicity of 90° but there are two energy minima, one for *cis* and one for *trans*. Often, the *cis*-configuration is less stable (*i.e.* has more energy) than the *trans*-configuration because the substituents 1 and 4 come close to each other in the *cis*-configuration. The highest energies are reached at torsion angles of 90° when the π-bond is broken and the planes defined by atoms 1–2–3 and 2–3–4 are perpendicular. C: Example of a molecule with two different substituents (black and magenta bonds), such as in butane or 1,2-dichloroethane. The larger magenta substituents (methyl groups in butane, chlorine atoms in 1,2-dichloroethane) give rise to local minima and maxima depending on how close they come. The torsion potential function is more complicated and has three cosine terms in this case. Note that torsion angles θ > 180° are usually given as θ–360, *e.g.* 240° is usually given as –120°, 300° equals –60°, etc.

17.2.2 Non-covalent interactions

Non-covalent or non-bonded interactions are through-space interactions between atoms that are not directly connected to each other by chemical bonds. Such interactions include electrostatic interactions of ions, permanent dipoles, or induced dipoles.

All non-covalent interactions are strongly distance-dependent, but to different extents (Table 17.2). The attractive terms of the energy potential functions have the general form $1/r^n$, where n is an integer between one and six, and r is the distance between the interacting atoms (Figure 17.6). The Lennard-Jones potential also takes into account the repulsive term, described by r^{12} (eq. 17.18).

TABLE 17.2:
Non-covalent interactions, sorted by strength and including distance (r) dependence.

type of interaction	distance dependence	energy range (kJ mol^{-1})	distance range (pm)	example
charge-charge	$1/r$	12–40	200–600	Arg···Asp
charge-dipole	$1/r^2$	< 30	320–450	Lys···O=C
dipole-dipole (incl. H-bond, CH-π)	$1/r^3$	4–30	240–320	NH···O=C
charge-induced dipole (cation-π)	$1/r^4$	< 20	450–510	Arg···Phe
dispersion (2x induced dipole)	$1/r^6$	2–6	330–440	Met···Phe

For comparison, the energy range of a covalent bond is 200–800 kJ mol^{-1} and the thermal energy RT at 25°C is 2.5 kJ mol^{-1}. Arg, Asp, Lys, Met, and Phe refer to amino acid side chains (Table 18.1).

Interactions with large n are characterized by a steep potential, and the strength of the interaction decreases strongly with increasing distance r. Such interactions are termed *short-range*. By contrast, smaller n give rise to less steep potentials, and a smaller decrease in interaction energy with increasing r. These interactions are termed *long-range* interactions. In the following, we will see that ionic interactions are strong and long-range ($n = 1$), while van der Waals interactions are weak and short range ($n \geq 6$).

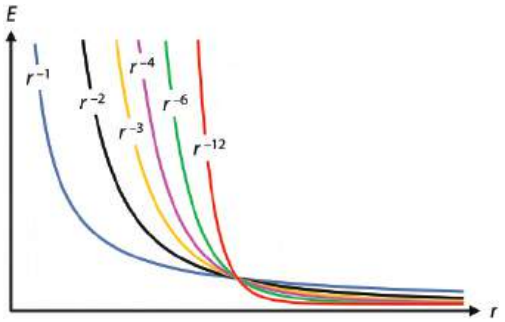


FIGURE 17.6: Graphs of the form $1/r^n$. Ionic interactions follow a potential with the least distance dependence ($n = 1$). At the other extreme with very strong distance dependence are the van der Waals interactions. The potentials with $n = 6$ and $n = 12$ are part of the van der Waals potential (Section 17.2.2.4). Curves with $n = 1…6$ describe attractions and that with $n = 12$ corresponds to a repulsion.

The different kinds of interactions between atoms and groups of atoms are typically divided into polar and non-polar interactions. Polar interactions involve ions ("*monopoles*") or *permanent dipoles* while non-polar interactions occur between neutral atoms and involve *induced dipoles*. Polar interactions are generally stronger than non-polar interactions. Both polar and non-polar interactions are affected by the dielectric constant ε_r of the medium and by the relative orientation of the dipoles involved.

17.2.2.1 Ionic interactions

The energy potential E_{ionic} for the pairwise interaction of charges at a distance r is described by Coulomb's law:

$$E_{ionic} = \frac{Z_1 \cdot Z_2 \cdot e^2}{4\pi\varepsilon_0\varepsilon_r \cdot r} \qquad \text{eq. 17.8}$$

Z_i is the number of charges of the interacting ions (including the sign of the charge), e is the magnitude of the elementary charge ($1.602 \cdot 10^{-19}$ C), ε_0 is the dielectric constant of vacuum ($8.85 \cdot 10^{-12}$ C V^{-1} m^{-1}), and ε_r is the *relative dielectric constant* of the medium in which the ionic interaction takes place. The product $4\pi\varepsilon_0\varepsilon_r$ is often summarized as the dielectric constant D. The value of ε_r is unity in vacuum, two for paraffin oil, 33 for methanol, and 81 for water. ε_r enters eq. 17.8 in the denominator, which means any value larger than unity leads to a weakening of the ionic interaction compared to vacuum. Solvents with high ε_r thus have a large damping effect on ionic interactions. For macromolecules in aqueous solutions, ionic interactions are weak on the surface and strong when the charges are buried in the non-polar interior of the macromolecule. The hydrophobic interior of proteins is often described by values of ε_r between 1 and 20. The smaller range of ε_r indicates that ionic interactions shielded from water within the hydrophobic core of a protein are 3–80 times stronger compared to the same interaction in water.

The dielectric constant D and thus the relative dielectric constant ε_r also occur in the denominator of the energy functions for all other types of polar interactions. Therefore, all polar interactions can be described by similar potential functions of the form $1/(D \cdot r^n)$. Ions are unique because their isolated charge is roughly spherically distributed over their surface, making them electrical monopoles. This means that the interaction between ions is independent of orientation, it is *isotropic*, and their interaction energy depends only on distance. Cations (Section 18.1.4.6) play an important role as structural parts of proteins (Ca^{2+}, K^+, zinc fingers), as prosthetic groups (Fe^{2+} in the heme of hemoglobin, Mg^{2+} in chlorophyll), in hydrolysis (Zn^{2+} in proteases and phosphatases, Ni^{2+} in urease), and in redox catalysis (iron-sulfur clusters, Mo-Fe clusters, Fe^{2+}/Fe^{3+} in the heme of cytochromes). Ubiquitous cations such as Na^+ and Mg^{2+} are part of the normal physiological environment and loosely bind to the surface of proteins, nucleic acids, and membranes. All cations bind to the lone pairs of protein residues, nucleic acids, prosthetic groups, counter anions, and/or water. Anions such as HPO_4^{2-}, HCO_3^-, and metabolites serve as physiological buffers and play roles in regulating enzyme activity.

Most polar interactions in biology, however, involve *dipoles*. As we will see in the following Sections, dipolar interactions depend not only on distance but in addition on the relative orientation of the interacting partners.

17.2.2.2 Interactions between ions and dipoles

A dipole is generated when the electron distribution around an atom or a group of atoms is asymmetric, giving rise to charge separation. In zwitter ions these charges can be integer multiples of the elementary charge, but *partial* charges are more common (δ^+ for $0 < \delta^+ < 1$ and δ^- for $-1 > \delta^- > 0$; Table 17.3). Dipoles can be described as vectors with a direction and a magnitude. The overall magnitude of the dipole depends on the magnitude of the charges q and the distance vector l between them (Figure 17.7). If the charges are of the same magnitude, the *dipole moment* μ is defined as:

$$\vec{\mu} = \mu = \boldsymbol{l} \cdot q \qquad \text{eq. 17.9}$$

In eq. 17.9 the dipole moment μ and the distance vector $\boldsymbol{l}$ are written in boldface, but without the vector symbol. The SI unit of the dipole moment is 1 C·m but the older unit Debye (1 D = $3.3 \cdot 10^{-30}$ C·m) is still in use.

TABLE 17.3:
Partial charges of some atoms in proteins.

residue	atom	partial charge
backbone	N	–0.36
	H_N	+0.18
	Cα	+0.06
	Hα*	+0.02
	C (carbonyl)	+0.45
	O (carbonyl)	–0.38
Ser / Thr	Cβ	+0.13 / +0.16
	Hβ	+0.02 / +0.01
	Oγ	–0.31
	Hγ	+0.17
Tyr	Oη	–0.33
	Hη	+0.16
Cys	Cβ	–0.105
	Hβ	+0.05
	Sγ	+0.01
	Hγ	+0.01
Lys / Lys^+	Nζ	–0.39 / –0.32
	Hζ	+0.15 / +0.32
Glu / Glu^-	$Oε_1$	–0.36 / –0.57
	$Oε_2$	–0.35 / –0.57
	$Hε_2$	+0.21 / n.a.

Data cited from Momany *et al.* (1975). J. Phys. Chem. 79(22): 2361–2381, with permission. * Pro and Gly have higher charges of about 0.04 and 0.05, respectively.

An overall neutral molecule may have a distribution of more than two partial charges, which sum to an overall dipole moment:

$$\mu_{tot} = \sum_i |\boldsymbol{l}_i| \cdot q_i \qquad \text{eq. 17.10}$$

where q_i and $|\boldsymbol{l}_i|$ are the magnitudes of the charges and their locations. For non-neutral molecules, with unequal partial charges, the resulting dipole moment depends on the location of the partial charges within the molecule:

$$\mu_{tot} = \sum_i |\boldsymbol{l}_i - \boldsymbol{l}_c| \cdot q_i \qquad \text{eq. 17.11}$$

with $\boldsymbol{l}_c$ usually taken as the center of mass of the molecule or assembly of particles.

The charges of a dipole can interact with neighboring ions and dipoles. The potential energy of the ion-dipole interaction (Figure 17.7) depends on the charges involved and the orientation of the interaction partners, and can be expressed by eq. 17.12.

$$E_{\text{ion-dipole}} = \frac{Z \cdot e \cdot (\boldsymbol{\mu} \cdot \boldsymbol{l})}{D \cdot |\boldsymbol{r}|^2} \qquad \text{eq. 17.12}$$

The dot product $(\boldsymbol{\mu} \cdot \boldsymbol{l}) = |\boldsymbol{\mu}| \cdot |\boldsymbol{l}| \cdot \cos\theta$ (Section 31.9.1) takes into account the angle θ between two distance vectors: one vector for the dipole and one vector for the distance between ion and the center of the dipole moment. The ion-dipole interaction follows a $1/r^2$ potential and for a given distance r is thus weaker than pure ionic interactions. The dielectric constant ε_r of the medium is again a damping factor for the interaction.

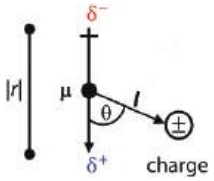


FIGURE 17.7: Charge-dipole interaction. A dipole with two partial charges (δ^+ and δ^-) separated by a distance vector ***l*** has a dipole moment **μ**. The extent of the interaction of such a dipole with a charge at position ***r*** is dependent on their relative orientation, expressed by the dot product (**μ**·***l***), and the sign of the charge. The direction of the dipole is from δ^- to δ^+.

What happens when we place two permanent dipoles close enough together so they can interact? In this case, both the distance between the dipoles and their relative orientation determine whether an attraction or repulsion results (Figure 17.8).

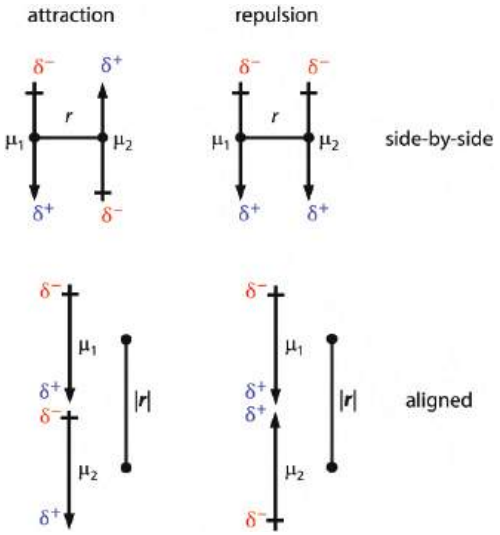


FIGURE 17.8: Interactions between two permanent dipoles. The top panel shows antiparallel and parallel side-by-side placements of dipoles. The bottom panel shows head-to-tail (parallel) and head-to-head (antiparallel) alignment of the dipoles. Placing the same polarities in proximity leads to repulsion, otherwise the dipoles attract each other.

The energy of the interaction between two dipoles is given by eq. 17.13.

$$E_{\text{dipole-dipole}} = \frac{\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2}{D \cdot |\boldsymbol{r}|^3} - \frac{3 \cdot (\boldsymbol{\mu}_1 \cdot \boldsymbol{r}) \cdot (\boldsymbol{\mu}_2 \cdot \boldsymbol{r})}{D \cdot |\boldsymbol{r}|^5} \qquad \text{eq. 17.13}$$

Again, the interaction is dependent on the dielectric constant D of the medium. The potential for the interaction between two permanent dipoles follows a $1/r^3$ dependence, and dipole-dipole interactions are thus weaker and of shorter range than interactions between dipoles and ions. The dot products ($\boldsymbol{\mu}_i \cdot \boldsymbol{r}$) describe the orientation of the dipole moments relative to the distance vector ***r*** (Section 31.9.1). In the special case of side-by-side orientation of dipoles, the dipole moments and the distance vector are perpendicular to each other, and $\boldsymbol{\mu}_i \cdot \boldsymbol{r} = 0$, so that eq. 17.13 reduces to

$$E_{\text{dipole-dipole}} = \frac{\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2}{D \cdot |\boldsymbol{r}|^3} \qquad \text{eq. 17.14}$$

For head-to-head or head-to-tail alignment of the dipolar and distance vectors, the dot products $\boldsymbol{\mu}_i \cdot \boldsymbol{r} = |\boldsymbol{\mu}_i \cdot \boldsymbol{r}|$, and eq. 17.13 simplifies to

$$E_{\text{dipole-dipole}} = \pm 2 \cdot \frac{\boldsymbol{\mu}_1 \cdot \boldsymbol{\mu}_2}{D \cdot |\boldsymbol{r}|^3} \qquad \text{eq. 17.15}$$

Here, the negative sign corresponds to head-to-tail alignment of the dipoles, which leads to attraction and thus a negative potential. We will see that in α-helices the dipole moments of the peptide bonds are aligned head-to-tail, summing to a permanent *helix dipole* (Figure 18.22).

The presence of an ion can also induce a dipole in an otherwise non-polar molecule. The interaction of an ion with such an induced dipole has a distance dependence of $1/r^4$ (eq. 17.16), which makes this interaction quite short range:

$$E_{\text{ion-induced dipole}} = \frac{Z \cdot e^2 \cdot \alpha}{2 \cdot D^2 \cdot |\boldsymbol{r}|^4} \qquad \text{eq. 17.16}$$

Z is the number of charges of the ion that induces the dipole in another molecule or molecular group, α is the *molecular polarizability* of the molecular group that generates the induced dipole under the influence of the ion. An example of an ion-induced dipole interaction that is frequently observed in biological molecules is the *cation-π interaction* (Figure 17.9), an interaction between positive charges and the π-electrons of an aromatic system. Cation-π interactions are formed between positively charged amino acids and aromatic protein side chains (Chapter 18) or nucleobases (Section 19.1). The positive charge can also be a metal ion, often Ca^{2+} or K^+, *e.g.* in DNA quadruplexes (Figure 19.20).

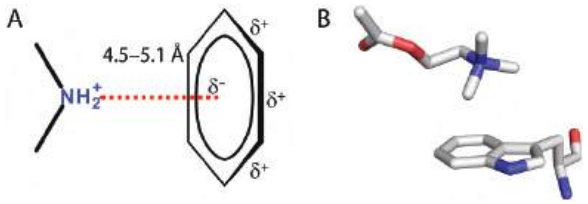


FIGURE 17.9: Cation-π interaction. A: The positively charged nitrogen of a secondary amine is centered on an aromatic plane, polarizing the π-electrons. Such interactions are frequently observed between the positively charged side chains of arginine and lysine and the aromatic side chains of phenylalanine, tyrosine, and tryptophan. Bases in nucleic acids can interact with charges in the same manner. B: An example of an enzyme/substrate cation-π interaction. The permanent positive charge of the quaternary ammonium ion acetylcholine (top) interacts with the indole side chain of tryptophan in acetylcholinesterase (PDB-ID 2ha4).

17.2.2.3 Hydrogen bonds

In Section 17.2.2.2, we discussed the interaction of dipoles and charges with other dipoles. While purely ionic interactions and interactions between ions and dipoles show little dependence on direction, the interactions between two dipoles are strongly dependent on their relative orientation.

The most important dipolar interaction in biology is the *hydrogen bond*. To understand hydrogen bonds, we first need to revisit the concept of electronegativity, polarized bonds, and partial charges. Electronegativity is a relative number between ≈ 1 and 4 that describes the tendency of an atom to attract electrons relative to other atoms (Table 17.4). Covalent bonds between two atoms with different electronegativity are polarized, resulting in a dipole with the negative end at the electronegative atom, and the positive end at the electropositive atom. Hydrogen has a low electronegativity of 2.2. When bound to an electronegative atom X (O, N, sometimes S and C in biological molecules), it will carry a partial positive charge due to shifting of the X-H bond electrons toward the atom X. In water this partial charge is about +0.4 on each hydrogen atom, leaving –0.8 on the oxygen atom. The larger the electronegativity of X, the larger the partial charges, and the larger the dipole moment of the

X-H bond. Hydrogen bonds are formed by precise alignment of two polarized bonds, where one of these bonds has a hydrogen atom at its positive end. The dipole carrying the hydrogen atom with the partial positive charge acts as the hydrogen bond *donor*. The partial positive charge of the hydrogen atom interacts with the negative end of an *acceptor* dipole. The resulting arrangement is X-H···X, with the hydrogen shared between two partners X with partial negative charges. The sharing of the hydrogen atom between the two electronegative atoms from donor and acceptor leads to appreciable orbital overlap: the distance between the hydrogen atom and the acceptor atom is shorter than the sum of the atomic radii. In other words, the hydrogen bond has partial covalent character. The partial covalent character sets the hydrogen bond apart from all other dipolar interactions and makes it one of the strongest non-bonded interactions in biology.

TABLE 17.4:
Electronegativity of atoms composing biological molecules.

non-metals	electronegativity	metals	electronegativity
H	2.2	Na	0.9
Li	1.0	Mg	1.3
C	2.6	K	0.8
N	3.0	Ca	1.0
O	3.4	Mn	1.6
F	4.0	Fe	1.8
P	2.2	Co	1.9
S	2.6	Ni	1.9
Cl	3.2	Cu	1.9
Br	3.0	Zn	1.7
I	2.7	Se	2.5

The numbers are rounded values of the Pauling electronegativity scale.

While a full understanding of the hydrogen bond requires quantum-mechanical treatment, some simple geometric considerations are sufficient to judge the energy of a hydrogen bond. The two main factors influencing the strength of a hydrogen bond are the magnitudes of the two dipoles involved and their relative orientation. The strongest hydrogen bonds are formed between dipoles that contain the electronegative oxygen and nitrogen atoms. Hydrogen bonds with the less electronegative sulfur or carbon atoms are also observed in macromolecules but usually do not contribute much to the overall stabilization energy. The strength of the interacting dipoles also influences the length of hydrogen bonds. For OH or NH as a donor and a carbonyl group as the acceptor, the distance between the electronegative atoms ranges between 270–320 pm (Figure 17.11). Shorter hydrogen bonds (by ca. 10 pm) are formed between permanent charges, *e.g.* between protonated amines and deprotonated carboxylic acids. Somewhat longer hydrogen bonds up to 320 pm are formed if the acceptor is sulfur or an imino-nitrogen.

The energetically most favorable orientation, on the other hand, is achieved when the hydrogen atom of the donor interacts with a lone pair of the acceptor in the direction of the acceptor molecular orbital. Thus, the three atoms including the central shared hydrogen atom plus the molecular orbital should be collinear with an angle of ideally 180°. Geometries with angles < 150° are not classified as hydrogen bonds. The four atoms involved in the hydrogen bond – two from the donor and two from the acceptor – should be coplanar for optimal dipole alignment (Figure 17.10).

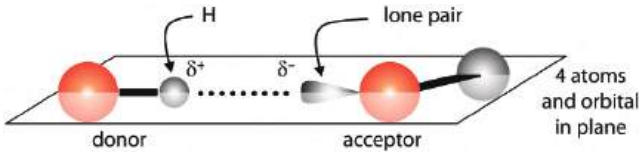


FIGURE 17.10: Hydrogen bond geometry. An electronegative oxygen atom is shown as a red sphere. Hydrogen and carbon are shown as gray spheres. In the donor dipole, the hydrogen atom carries the positive partial charge δ^+. The negative partial charge δ^- of the acceptor dipole is concentrated in the lone pairs. One of them is shown as a gray lobe. The hydrogen bond is shown as a dashed line. The optimal geometry of a hydrogen bond is in the direction of the lone pair, with the four atoms O-H…O-C in the same plane.

The energy range of hydrogen bonds is large – they range from just 5 kJ mol^{-1} to 40 kJ mol^{-1} *in vacuo*. These energies are reduced in solvents of increasing relative dielectric constant. Hydrogen bonds in proteins typically have energies of 2–6 kJ·mol^{-1}. Similar to any electrostatic interaction, the energy depends on the dielectric constant ε_r of the solvent. Hydrogen bonds exposed to water are weaker than hydrogen bonds that are shielded from water, for example in the hydrophobic interior of proteins. High-resolution crystal structures show that in the core of proteins, most of the hydrogen bond donors and almost all of the hydrogen bond acceptors are saturated, indicating that it is energetically unfavorable to leave hydrogen-bonding valences unoccupied.

Water not only reduces the strength of hydrogen bonds due to its high relative dielectric constant ε_r, but also is both a powerful hydrogen bond donor and acceptor by itself. With the exception of membranes, water is the solvent for most biological systems, and therefore effectively competes for hydrogen bond formation with the solute. Thus, a hydrogen bond at the surface of a protein has little energetic contribution to stability because it is almost isoenergetic to a hydrogen bond formed with water. By contrast, hydrogen bonds buried in the interior of macromolecules experience no competition from bulk solvent and are stronger due to the reduced relative dielectric constant ε_r.

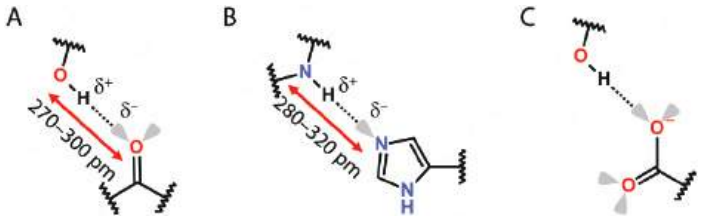


FIGURE 17.11: Examples of hydrogen bonds. The hydrogen bond is shown as a dashed line and orbitals are drawn as gray lobes. The distance between the hydrogen atom and the electronegative atom (N or O) is shorter than the sum of atomic radii, indicating attractive orbital overlap. Generally, the O-H···O or N-H···N angle is > 150°, with 180° being optimal. A: The hydroxyl group of an alcohol binds to a carbonyl group. B: An NH-group is the donor and the imino nitrogen of imidazole is the acceptor. C: Charged hydrogen bonds are slightly shorter and more stable than non-charged hydrogen bonds.

A network of several hydrogen bonds can have a strongly stabilizing effect. Hydrogen bond networks are cage-like structures between several donors and acceptors that are held together by mutual and reciprocal hydrogen bonds. Such networks are ubiquitous in biological macromolecules. The formation of the network is highly cooperative, and as a result it can be very sensitive to disturbances. It is sometimes sufficient to remove a single hydrogen bond by mutation of a side chain, dissociation of a ligand, or hydrolysis of a nucleotide in order to destroy an entire network of hydrogen bonds and drastically change the conformation of a protein. The cooperativity of hydrogen bond networks is used for the regulation of many cellular processes.

Hydrogen bonds are central to interactions in biology. Together with the hydrophobic effect (see Section 18.3.6.3), hydrogen bond formation is a major driving force in protein folding (Figure 17.11; Section 18.4.1). *Intramolecular hydrogen bonds*, with the hydrogen bond donor and acceptor groups within the same molecule, dictate the preferred conformation of molecules. *Intermolecular hydrogen bonds* between hydrogen bond donor and acceptor groups on different molecules, are essential for all kinds of molecular recognition events both between macromolecules and between macromolecules and ligands (Section 18.3.7). The strong orientation dependence of the hydrogen bond energy provides a strong geometric constraint that enables the read-out of hydrogen donor and acceptor patterns in interactions. The complementarity of these patterns not only contributes to binding affinity, but also is the key for binding specificity of biomolecular interactions.

A pronounced directionality for the strength of interactions is also observed in *halogen bonds* (Box 17.1), where chlorine, bromine, or iodine atoms accept lone pairs from other atoms. This interaction is not prominent in nature but has applications in inhibitor design.

BOX 17.1: HALOGEN BONDS.

The halogen bond is an attractive electrostatic interaction between a halogen atom and a lone pair from another atom (Figure 17.12). The interaction is highly directional, a feature it shares with the hydrogen bond (Figure 17.11) and that enables halogen bonds to establish selectivity for ligand binding. Halogen atoms Cl, Br, and I (but not F) in organic molecules have anisotropic charge distribution: the three lone pairs form a central belt of negative charge around the halogen atom and isolate a region of positive electrostatic potential on the outermost tip of the halogen atom's surface that is farthest away from other atoms (Figure 17.12). This positively polarized region is termed the σ-*hole* and acts as an acceptor in the halogen bond. The strength of a halogen bond correlates with the size of the σ-hole and the polarizability of the halogen. Electron-withdrawing substituents bound to the halogen increase the σ-hole: halogen bonding becomes therefore stronger in the sequence Cl < Br < I.

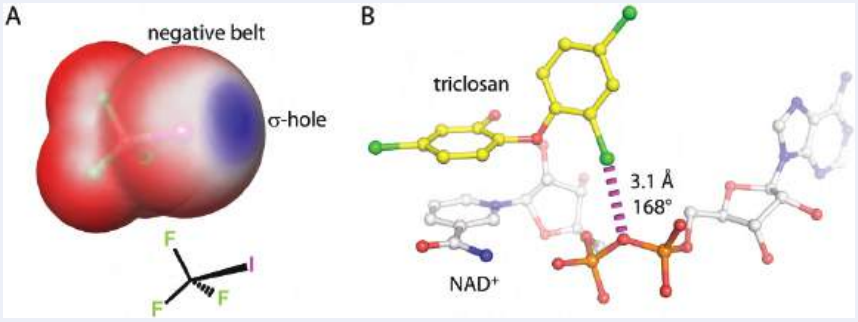


FIGURE 17.12: Halogen bonding. A: The molecular surface of trifluoromethyl iodide, CF_3I, is colored according to negative (red) and positive (blue) electrostatic potential. The electronegative fluorine atoms pull the electrons of the carbon-iodine σ-bond toward them, resulting in a σ-hole at the extension of the C-I bond (visible as a blue patch). B: Triclosan, a pesticide, bound to the NAD^+ cofactor in enoyl-acyl carrier protein reductase. One of the chlorine atoms (green) forms a halogen bond (magenta dashes) with an oxygen atom of NAD^+ (PDB-ID 3pjf).

As in hydrogen bonds, the optimal angle for halogen bonding between donor (the C-Hal bond) and the acceptor (the atom carrying a lone pair) is 180°, but smaller values down to 140° have been observed. The distance between the donor and acceptor atom is slightly shorter (ca. 7%) than the sum of their van der Waals radii, and the energy of a halogen bond is on the order of 2–8 kJ mol^{-1}. Only a few hundred halogen bonds have been identified in biological systems, but the use of halogen bonds is becoming popular in structure-based drug design.

17.2.2.4 Interactions between induced dipoles: van der Waals interactions

Even molecules without a permanent dipole will attract each other electrostatically. The reason is the fluctuation of the electron distribution around an atom, which leads to a transient dipole. This transient dipole polarizes electrons from neighboring molecules, thereby inducing another transient dipole in these neighbors. Although the induced dipoles are of tiny magnitude and very short-lived, they lead to a weak attraction between the molecules. The attraction depends on the inverse 6th power of the distance, making it a very short-range interaction compared to the other dipolar interactions (Table 17.2). The attractive force between two induced dipoles has been predicted by quantum mechanics and is also known as the *London dispersion force*. The dispersion potential E_{London} depends on the distance r, the dispersion coefficient C_6 that is a measure for the strength of the interaction, and the polarizabilities α_i of the interacting atoms:

$$E_{\text{London}} = -\frac{3C_6\alpha_1\alpha_2}{4r^6} \qquad \text{eq. 17.17}$$

The polarizability is a property of an individual atom or of a group of atoms that depends on the volume and the number of electrons contained in that volume (the electron density). The polarizability is inversely proportional to the relative dielectric constant ε_r of the surrounding medium, with the effect that London dispersion forces are weaker in water compared to hydrophobic environments. One of the most important driving forces for protein folding is the burial of hydrophobic amino acid side chains in the interior of proteins, away from the solvent. The phenomenon is known as the hydrophobic effect (Section 18.3.6.3). The individual London dispersion forces are weak, but there are many of them in a single molecule and they sum to large contributions to its overall energy. London dispersion forces are responsible for much of the stability of biological molecules and the binding energies between them.

The dispersion potential (eq. 17.17) predicts a stronger attraction the closer the atoms come together. However, below a certain distance the electrostatic repulsion of the electron clouds and, at even closer distance, the repulsion between the positively charged nuclei sets in. The strong repulsive force between atoms at close range is described by an additional energy term that depends on the inverse 12th power of the distance. This energy term modifies the London energy to what is known as the van der Waals energy E_{vdW}:

$$E_{\text{vdW}} = \frac{A}{r^{12}} - \frac{B}{r^6} \qquad \text{eq. 17.18}$$

The potential function in eq. 17.18 is the *Lennard-Jones potential* or *6–12 potential*, describing the van der Waals energy E_{vdW}. The actual values for the repulsive coefficient A and the attractive coefficient B depend on the type of interacting atoms and are tabulated. The balance of the attractive London dispersion forces and the repulsive forces of the nuclei produces an energy minimum at a distance r_0 termed the *van der Waals contact distance* (Figure 17.13). At this distance, the net attractive force between the atoms is maximal.

The van der Waals contact distance is the sum of the van der Waals radii of the interacting atoms. Van der Waals radii of atoms and groups of atoms that are important in biology are given in Table 17.5.

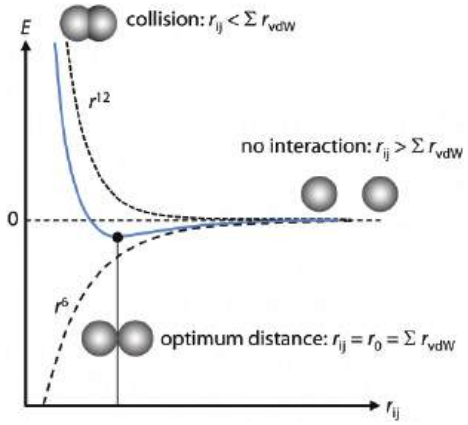


FIGURE 17.13: Lennard-Jones potential. Atoms *i* and *j* are at minimum energy at the van der Waals contact distance r_0. The shape of the Lennard-Jones potential also applies to hydrogen bonds (Section 17.2.2.3).

TABLE 17.5:
Van der Waals radii (in pm) of selected atoms and groups.

H 120	C 170–177	N 155–166	O 150–152	F 146–147
	CH_2 180	P 180–190	S 180–189	Cl 175–182
	CH_3 185		Se 182–190	Br 183–186
				I 198–204

Radii are mean values from small molecules in the Cambridge Structural Database.

Van der Waals interactions between the aliphatic chains of lipids drive the formation of membranes (Section 18.4), and those between hydrophobic amino acids, both aliphatic and aromatic, drive the formation of the hydrophobic core in proteins and contribute substantially to protein stability (Section 18.5). Aromatic groups present in nucleic acids (Figure 19.2), many prosthetic groups (Figure 21.13), and small molecules are also frequently involved in van der Waals interactions (Figure 17.14). The large van der Waals radius and the comparatively large polarizability of sulfur atoms make sulfur a good binding partner for aliphatic and aromatic systems. Cysteine and methionine are often found in protein cores where they form contacts of about 350–400 pm mean distance.

The energies of der Waals interactions show little dependence on orientation, but in some cases, certain arrangements are energetically more favorable. For example, in interactions between aromatic systems, two geometries termed face-to-face and edge-to-face orientations are preferred (Figure 17.14). The face-to-face interaction is characteristic for the base stacking of nucleic acids (Section 19.4.1), and the edge-to-face orientation is found between aromatic groups in protein-DNA interactions. The π-systems of the amide bonds in peptides and proteins can also favorably stack on aromatic systems.

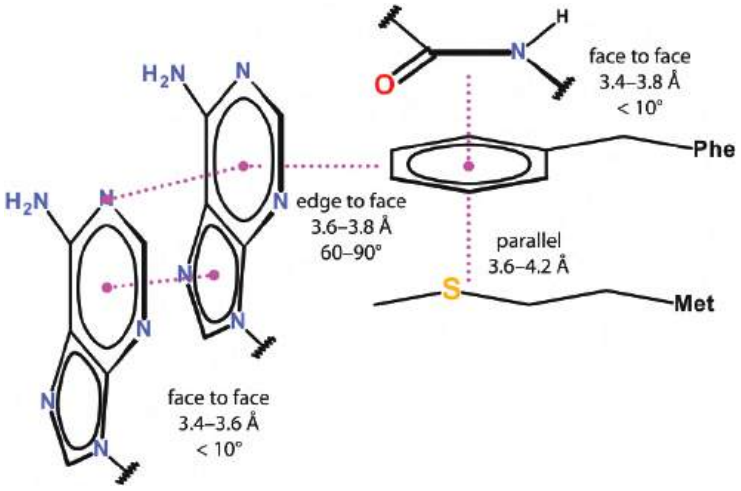


FIGURE 17.14: Face-to-face and edge-to-face interactions. Interactions are illustrated for an amide bond (top), the aliphatic methylene groups of a methionine side chain, and aromatic groups such as the nucleobase adenine and the side chain of phenylalanine. Amide groups stack almost parallel on top of aromatic systems with a tilt of generally less than 10°. Aromatic groups tend to pack either perpendicular (edge-to-face) or parallel to each other (face-to-face). Often, the planar groups are shifted relative to each other in the parallel orientation.

17.3 ENERGY LANDSCAPES AND THE MOST STABLE CONFORMATION

We have seen in Section 17.2 how we can describe the energy of any conformation of an isolated molecule as the sum of its covalent and non-covalent interactions. These energy terms depend on the arrangement of the atoms in space, which is defined by the coordinates of all atoms in the molecule. If we know the coordinates of all atoms in a molecule, we can in principle calculate its energy. If we do so for all possible conformations of a molecule, we can generate a multidimensional plot of all energies as a function of the conformation, which constitutes the *energy surface* of the molecule. Minima in the energy surface represent thermodynamically stable conformations, maxima are transition states on the path between different conformations.

Energy surfaces have been constructed from observed conformational preferences of macromolecules. The energy of the protein main-chain is represented by the Ramachandran plot (Figure 18.21), the side-chains have preferred torsion angles (Figure 18.5 in Section 18.1.3) and proteins fold along pathways that agree well with calculated energy surfaces (Section 18.5.3). The ribose-phosphate backbone of nucleic acids also adopts minima in its own energy surface (Figure 19.12 in Section 19.3.2).

To calculate an energy surface of all conformations for a molecule, we would need to know the energy of every single conformation possible. To compute these energies, we could use a *divide et impera* approach: Divide the molecule into small groups of atoms and calculate the pairwise interaction energies using tabulated reference values. The overall energy of the molecule would then be obtained by summing these pairwise interaction energies. This procedure, reminiscent of the definition of the standard free energy of formation (Section 1.6) if the division is into single atoms, must remain an approximation, as the energy of pairwise interactions is context-dependent.

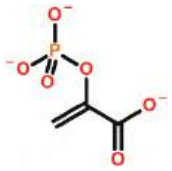
A more practical approach is quantum mechanical *ab initio* calculations. Often done *in vacuo* to simplify the system, these calculations try to estimate the total energy of a molecule based on its calculated electron distributions without inclusion of experimental data. Other methods use empirical *energy potentials* that describe the lowest energy, or ground state of a molecule as a function of distances and angles between atoms (Section 20.1). The energy terms in such functions are derived from the conformational preferences observed in highly precise and accurate crystal

structures of small organic and inorganic molecules (collected in the Cambridge Structural Database). The currently best method in calculating the overall lowest energy and most probable conformer of small molecules is to combine quantum mechanical and energy potential calculations. For reasons of simplicity the calculations are often done *in vacuo* or with just a few solvent molecules surrounding the target molecule. Although the calculations are dominated by enthalpic contributions caused by the configuration, and tend to neglect entropic contributions, the resulting lowest-energy conformations are usually accurate.

In the following Chapters 18 and 19, we will turn our attention to the structure of proteins, membranes, and nucleic acids. We will see how the different types of interactions discussed in this Chapter enable inter- and intramolecular interactions, and will revisit energy surfaces of protein (Figure 18.21, Section 18.5) and nucleic acid conformations (Figure 19.12).

QUESTIONS

17.1 The hydrolysis of phosphoenol pyruvate (shown) to pyruvate and inorganic phosphate is thermodynamically strongly favored, liberating ca. –62 kJ mol^{-1}. What could be the reason?



17.2 The C-N bond in amide groups has partial double bond character. Can this be explained by tautomerism?

17.3 In 1953, Robert Corey and Linus Pauling proposed an approximate relationship for the strain energy E (in kcal·mol^{-1}) associated with rotation about the C-N bond in a peptide plane as a function of the torsion angle ω:

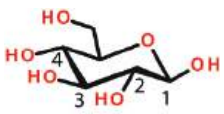
$$E = 30 \cdot \sin^2 \omega$$

What is the amount of energy required for ± 5° deviation from planarity (0°)? How does it compare to the average thermal energy of an atom in an ideal gas of $1.5 \cdot k_B \cdot T$ at room temperature? Why are deviations of > 15° so rarely seen in protein structures? What is the activation energy predicted for the *trans*-to-*cis* isomerization and how does it compare to the measured *cis*-to-*trans* isomerization of 14–24 kcal·mol^{-1}?

17.4 About 5-6% of Xaa-Pro peptide bonds in proteins are in the *cis*-configuration, where Xaa is any amino acid residue. What are the "statistical" equilibrium constants for 5% and 6% occurrence and what are the associated free energies at 25°C and 37°C? What are some limits for interpreting statistical occurrences as equilibria?

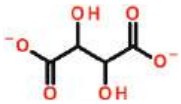
17.5 Methionine sulfoxide is a natural oxidation product in antibodies. Is this side chain chiral?

17.6 The structure of glucose is shown. Draw the structures of mannose (C2-epimer) and galactose (C4-epimer).



glucose

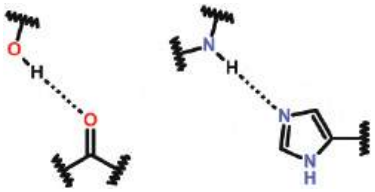
17.7 The structure of tartrate is shown. How many stereocenters and how many stereoisomers does this ion have?



17.8 For the potential functions of the form $1/r^n$, what is the change in energy (in %) when the distance r is increased by 10% for $n = 1$ and $n = 6$?

17.9 The short-range effect of the $1/r^n$ potentials can be visualized by calculation of the distance change that is necessary to halve the interaction energy. Derive the general formula for the factor by which the distance has to change in order to halve the energy as a function of n.

17.10 For the hydrogen bonds shown, assign the possible protein side chains and main-chain parts that can engage in them as donor or acceptor.



17.11 N-methyl acetamide dimerizes in solution by forming hydrogen bonds. Draw the structure of the dimer. The dimerization is favored in non-polar solvents such as CCl_4, but disfavored in water. What is the reason?

REFERENCES

Alvarez, S. (2013) A cartography of the van der Waals territories. *Dalton Trans* **42**(24): 8617–8636.
· statistical analysis of van der Waals contacts in molecules

Bissantz, C., Kuhn, B. and Stahl, M. (2010) A medicinal chemist's guide to molecular interactions. *J Med Chem* **53**(14): 5061–5084.
· discusses the preferred conformations of small molecules

Clark, T., Hennemann, M., Murray, J. S. and Politzer, P. (2007) Halogen bonding: The sigma-hole. Proceedings of "Modeling interactions in biomolecules II", Prague, September 5th–9th, 2005. *J Mol Model* **13**(2): 291–296.
· detailed explanation on how the σ-hole is formed

Desiraju, G. R. and Steiner, T. (2006) *The Weak Hydrogen Bond in Structural Chemistry and Biology*. Oxford University Press.
· in-depth treatment of the hydrogen bond

Hardegger, L. A., Kuhn, B., Spinnler, B., Anselm, L., Ecabert, R., …, Diederich, F. (2011) Systematic investigation of halogen bonding in protein-ligand interactions. *Angew Chem Int Ed* **50**(1): 314–318.
· role of halogen bonds in structure-based drug design.

Hassel, O. and Hvoslef, J. (1954) The structure of bromine 1,4-dioxanate. *Acta Chem Scand* **8**(5): 873.
· first description of a halogen bond

Klostermeier, D. and Millar, D. P. (2002) Energetics of hydrogen bond networks in RNA: Hydrogen bonds surrounding G+1 and U42 are the major determinants for the tertiary structure stability of the hairpin ribozyme. *Biochemistry* **41**(48): 14095–14102.
· an example of the sensitivity of hydrogen bond networks, analyzed by double- and triple-mutant cycles

Scholfield, M. R., Zanden, C. M., Carter, M. and Ho, P. S. (2013) Halogen bonding (X-bonding): A biological perspective. *Protein Sci* **22**(2): 139–152.
· review of halogen bonds in biology and their relation to hydrogen bonds

18 Proteins

The name protein originates in 1838 from the Greek πρώτειος (*proteios*, primary), indicating an essential substance for life. It was originally believed that one gene codes only for a single protein (*one-gene-one-polypeptide hypothesis*), but in the human genome only 20000–25000 genes code for > 100000 proteins that constitute our *proteome*, the sum of all proteins in our body. The observed protein variety is generated by several mechanisms acting on the level of DNA, messenger RNA (mRNA), and the proteins, including gene shuffling, alternative transcription and translation initiation and termination, mRNA splicing, post-translational modification, proteolytic maturation, and somatic mutagenesis in immune cells. While in the early 1900s, proteins were viewed as unstructured colloids akin to wallpaper paste, the first crystal structure of myoglobin in 1958 showed that proteins are well-defined structural entities. Modern methods for the determination of protein structures are discussed in Part IV (Chapters 22, 24, and Section 26.5) of this book.

We begin our treatment of protein structure by first introducing the building blocks, the amino acids, followed by how these condense into their polymers, the proteins (Sections 18.1 and 18.2), and then how proteins fold into their functional three-dimensional conformation (Section 18.3). In Section 18.4, we will take a similar approach for membrane assembly from lipids (Section 18.4.3), and connect the two areas in the discussion of membrane proteins. Finally, we will treat the principles governing protein folding and stability (Section 18.5).

18.1 AMINO ACIDS AND THE PEPTIDE BOND

18.1.1 Properties of the twenty canonical amino acids

Proteins are polymers of amino acids that are linked *via* amide bonds, also termed *peptide bonds* (Section 18.1.2). There are twenty *canonical* amino acids. According to the central dogma of molecular biology, the amino acid sequence of a protein is encoded in the DNA by nucleotide triplets, transcribed into mRNA, and assembled into proteins during translation by the ribosome. The diversity of proteins is further increased by post-translational chemical modification of the standard amino acids (phosphorylation, glycosylation, oxidation; Section 18.1.4).

The protein-forming twenty amino acids have a common structure with an ammonium group, a carboxylate group, and a side chain attached to a central carbon atom named Cα. The pK_A values of the NH_3^+-group range between 6.8–8.0, and those of the CO_2^--group are between 3.5–4.3.

At physiological pH, amino acids exist as zwitter ions, with a positively charged amino group and a negatively charged carboxyl group. When condensed into a polypeptide chain, only the N- and C-terminal residues retain a free NH_3^+- and CO_2^--group, respectively. The amino acids in the polypeptide are called *residues*.

Except for glycine, all canonical amino acids are chiral, with the same stereochemistry at Cα. In the (*R*/*S*) convention, it is the *(S)*-configuration that is observed in nature except for cysteine, which somewhat confusingly has to be called an *(R)*-configuration just because of a change in substituent priority (Section 17.1). Amino acids differ in their side-chain properties (Figure 18.1; Table 18.1), and these properties in turn define the structure and properties of proteins.

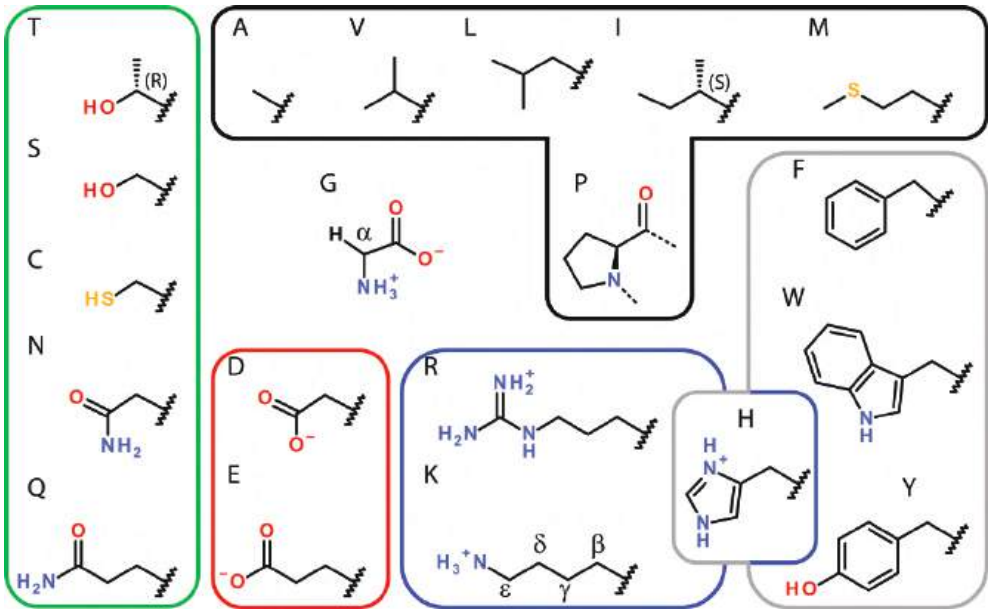


FIGURE 18.1: The twenty canonical amino acids. The structure of glycine is shown in its zwitterionic form; the Cα atom is labeled. Glycine has only a hydrogen atom as the side chain and thus is achiral. The other side chains are grouped according to chemical nature, hydrophobicity, and acidity. Labels are given in single letter code (Table 18.1). Functional groups are charged as they would be at pH 7. Aliphatic side chains are grouped in a black box, but note that the methylene groups of glutamine, glutamate, arginine, and lysine and the methyl group of threonine also have aliphatic properties. Aromatic side chains are grouped in a gray box. Histidine can be charged under physiological conditions, so it belongs to both the hydrophobic and the positively charged groups (blue box). The negatively charged aspartate and glutamate (red box) are placed next to their isosters asparagine and glutamine. Polar side chains are grouped in a green box. Threonine and isoleucine are the only amino acids with chiral side chains. Cysteine is redox active and can form disulfide bonds. Proline is the only amino acid where the side chain connects to the main chain. The resulting ring is a pyrrolidine.

Amino acids are typically abbreviated either by their three-letter or their single-letter codes. The single-letter code is useful for protein sequence alignments (Section 18.2.2), but the readability of text is better when using the three-letter code. In this book, both systems are used, depending on the context. Starting from Cα, the atoms in the side chains are usually labeled β, γ, δ, etc.

Amino acid side chains can be grouped according to their biophysical properties *i.e.* charged or uncharged, polar or non-polar, aliphatic or aromatic. There is overlap between the categories because many side chains have regions with different biophysical properties. For instance, the methylene groups of the polar Lys, Glu, and Gln frequently engage in van der Waals interactions at the same time when their terminal groups are involved in charge-charge or hydrogen bond interactions.

TABLE 18.1:
Properties of the twenty canonical amino acids.

name	3-letter	1-letter	pK_A side chain	residue mass (Da)	hydro-phobicity*	hydrogen bond	functionality
alanine	Ala	A		71.08	0.25 / 1.8	–	aliphatic
cysteine	Cys	C	8.6–9.5	103.14	0.04 / 2.5	weak D/A	polar, hydrophobic
aspartate	Asp	D	3.7–4.0	115.09	–0.72 / –3.5	A	charged
glutamate	Glu	E	4.2–4.5	129.11	–0.62 / –3.5	A	aliphatic, charged
phenylalanine	Phe	F		147.18	0.61 / 2.8	–	aromatic, hydrophobic
glycine	Gly	G		57.05	0.16 / –0.4	–	–
histidine	His	H	6.0–7.0	137.14	–0.40 / –3.2	D/A	aromatic, polar, charged positive or neutral
isoleucine	Ile	I		113.16	0.73 / 4.5	–	aliphatic
lysine	Lys	K	10.4–11.1	128.17	–1.10 / –3.9	D	aliphatic, charged positive
leucine	Leu	L		113.16	0.53 / 3.8	–	aliphatic
methionine	Met	M		131.20	0.26 / 1.9	–	aliphatic
asparagine	Asn	N		114.10	–0.64 / –3.5	D/A	carbonyl and $-NH_2$
proline	Pro	P		97.12	–0.07 / –1.6	–	aliphatic
glutamine	Gln	Q		128.13	–0.69 / –3.5	D/A	aliphatic, carbonyl and $-NH_2$
arginine	Arg	R	12.3–12.6	156.19	–1.76 / –4.5	D	aliphatic, charged positive
serine	Ser	S	~16	87.08	–0.26 / –0.8	D/A	polar, –OH
threonine	Thr	T	~16	101.10	–0.18 / –0.7	D/A	aliphatic, polar, –OH
valine	Val	V		99.13	0.54 / 4.2	–	aliphatic
tryptophan	Trp	W		186.21	0.37 / –0.9	D	aromatic, NH
tyrosine	Tyr	Y	9.6–10.3	163.17	0.02 / –1.3	D/A	aromatic, –OH

The pK_A of the N-terminal NH_3^+-group is 6.8–8.0, and that of the C-terminal carboxyl group is 3.5–4.3. All pK_A values can be shifted by several units due to the protein environment. The mass of the amino acid residue in a protein (in Da) is the molecular mass minus that of water. * The first number is from Eisenberg *et al.* (1982) Faraday Symp. Chem. Soc. 17(0): 109–120, and the second is from Kyte & Doolittle (1982) J. Mol. Biol 157(1): 105–132.

The NH-group of the aromatic Trp side chain, an indole, frequently acts as a hydrogen bond donor while the faces of the indole pack against other hydrophobic surfaces. The same combination of polar and non-polar interactions is frequently observed for the phenol side chain of Tyr and the imidazole side chain of His. The thiol group in Cys can engage in weak hydrogen bonds, in van der Waals interactions, and in metal binding. The properties of a side chain may also change as a function of pH, hydrogen-bonding environment, temperature, or dielectric constant of the environment. A good example is the aromatic imidazole side chain of His, which can be charged or neutral and, depending on its protonation state, act as hydrogen bond donor, acceptor, or both (Figure 18.2).

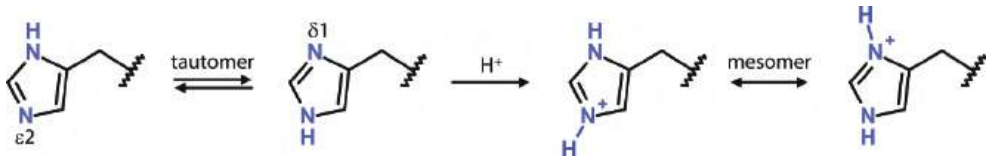


FIGURE 18.2: Protonation states of His. The tautomer with the hydrogen at Nε2 is slightly more stable than the other tautomer with the hydrogen at Nδ1. After protonation of the imidazole side chain, the positive charge is stabilized by resonance. Both mesomeric structures are shown (note the different arrow types for a chemical equilibrium and mesomeric structures).

Histidine is the only amino acid whose side-chain pK_A value is close to the physiological pH, facilitating protonation and deprotonation, and making it well-suited for acid-base catalysis. As a consequence, histidine is the most frequent residue observed in enzyme active sites (Section 16.2.1), although it is one of the less frequent amino acids in proteins overall.

Side-chain pK_A values depend on the local environment. Often, the pK_A values of side chains in proteins are substantially different from the value of the isolated amino acid. In many enzymes, specific local environments around the catalytic amino acid stabilize the protonated or deprotonated form of the side chain so it can be used in catalysis. This is true for the side chains of amino acids Cys, Ser, Thr, Tyr, and Glu. Proteases are a well-studied enzyme class with shifted pK_A values. In Cys-, Ser-, or Thr-proteases, the protein environment makes these side chains more acidic by stabilizing the deprotonated form, which increases their nucleophilicity for cleavage of amide bonds in proteins. Similarly, in protein disulfide isomerases, the protein environment reduces the pK_A of an active site Cys to about 4.5, four orders of magnitude from the normal. The stabilized thiolate anion acts as a nucleophile to accelerate disulfide exchange, a rate-limiting step in protein folding (Section 18.5.5.2).

18.1.2 The peptide bond

The amide group formed upon condensation of amino acids is known as a *peptide group* (Figure 18.3), and the bond connecting the amino acids is the *peptide bond*. The peptide bond has partial double bond character because of conjugation of the π-electrons. Due to the large electronegativity of oxygen, the π-electrons are shifted toward the oxygen atom, establishing a permanent dipole with its positive end at the NH hydrogen atom and the negative end at the oxygen atom of the CO-group. This dipole is especially important for hydrogen bonds that stabilize the secondary structure elements in proteins (Section 18.2).

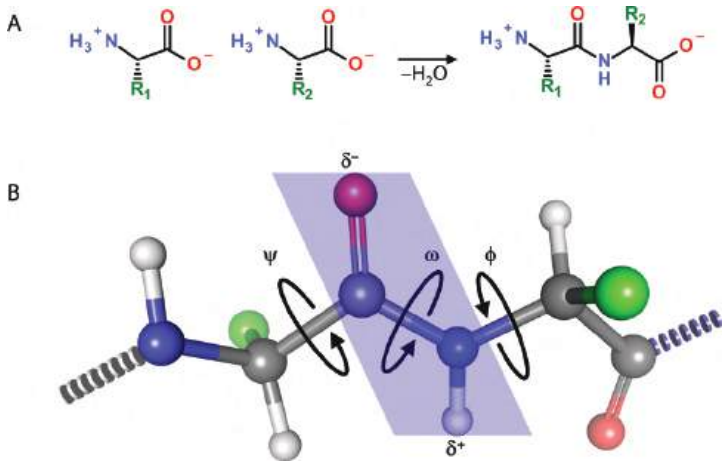


FIGURE 18.3: Formation and properties of the peptide bond. A: Formation of a dipeptide by condensation of the amino and carboxyl groups of amino acids. B: Overall structure of the polypeptide chain. The planar peptide group is highlighted (blue). The amide is a dipole with its negative part at the carbonyl oxygen (δ^-). The torsion angles ϕ, ω, and ψ are indicated. ϕ and ψ can adopt many values and define the conformation and flexibility of the polypeptide chain. ω is either 0° or 180°, and determines whether a peptide bond is *cis* or *trans*. Side chains are indicated as green spheres.

A polymer of amino acids is termed a *polypeptide* or just *peptide* when it is small, and *protein* when it is larger. The repeating sequence of atoms in a polypeptide, $-[\text{-N-C}\alpha\text{-C-}]_n-$, is the *main chain* or *backbone*. While the backbone conformation determines the secondary structure of the polypeptide, the amino acid *side chains* connected to the Cα atoms provide chemical and biophysical properties, and they stabilize the three-dimensional structure of the protein.

The structural context of the peptide bond is described by the torsion angles ψ and ϕ (Figure 18.3). ψ describes the rotation around the Cα-C bond and ϕ the rotation around the N-Cα bond. The torsion angles ψ and ϕ can adopt many different values, giving polypeptides considerable conformational flexibility. Not all combinations of ψ and ϕ are allowed, however, as some lead to steric clashes between atoms. Certain ψ and ϕ combinations are characteristic for the secondary structure of the main chain.

The torsion angle ω describes the rotation about the peptide bond (Figure 18.3). In contrast to ψ and ϕ, ω is much more restricted, and it clusters predominantly around two values: 0° and 180°, corresponding to the *cis*- and *trans*-configurations of a planar peptide group, respectively.

During ribosomal biosynthesis peptide bonds are formed in the *trans* configuration. Many proteins, however, require *cis*-peptide bonds for their function, which are introduced through post-translational isomerization. For *trans-cis* isomerization, the conjugation of the peptide bond needs to be broken. The total rotational energy barrier to remove the conjugation in the peptide bond is about 80 kJ mol^{-1}, indicating a very slow isomerization with time constants in the minute range. Rotation of the peptide bond by just 10° out of the plane already requires about 4 kJ mol^{-1}. Analysis of high-resolution protein crystal structures showed that in about 10–15% of all *trans*-peptide bonds, ω deviates from 180° by ≥ 10°, with occasional deviations of over 30° from planarity. So while in most cases the peptide bond is planar, some distortion of the *trans*-peptide plane is normal.

In *cis*-peptide bonds, the orientation of the CO- and NH-groups is inverted with respect to the *trans*-configuration. Compared to the *trans*-configuration, a *cis*-configuration introduces a sharp bend into the polypeptide chain, which allows for very tight turns (Section 18.3.3). The sizeable conformational change associated with *cis*/*trans* isomerization is used in biology as a molecular switch (Box 18.10).

In contrast to the high activation energy separating *trans* and *cis* isomers, their energetic difference is moderate: a *trans*-peptide is more stable than a *cis*-peptide by 2–10 kJ mol^{-1}, depending on which amino acids form the peptide bond. The main reason for the lower stability of *cis*-peptide bonds is a steric clash between the side chains in the *cis*-configuration (Figure 18.4).

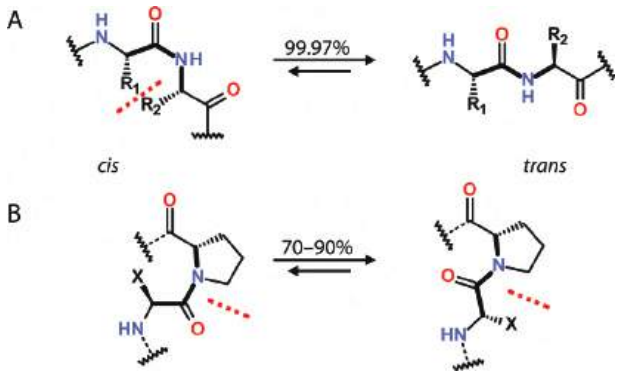


FIGURE 18.4: The *cis*-peptide bond. A: The side chains in a non-prolyl *cis*-peptide clash with each other (dashed red line) but point away from each other in the preferred *trans*-peptide. B: In a prolyl peptide, both *cis*- and *trans*-configurations lead to clashes, making the energy difference between these states smaller than in the non-prolyl peptide bonds. The amino acid X preceding the proline can in principle be any amino acid. The aromatic side chains His, Phe, Tyr, and Trp can favorably stack on top of the aliphatic pyrrolidine of Pro – they are more frequently found in this position than the other side-chains.

If the C-terminal residue of the peptide bond is Pro, the *cis*-configuration is only slightly less stable than the *trans*-configuration by about 2 kJ mol^{-1}. In fact, most *cis* peptide bonds in proteins are X-Pro bonds.

Why is the peptidyl-prolyl peptide bond a special case? In solution, 10–30% of all X-Pro sequences in peptides are in the *cis*-configuration, in agreement with the ratio expected from the ΔG values. In an X-Pro peptide bond, the pyrrolidine side chain clashes both in the *cis* and in the *trans* configurations (Figure 18.4), which is why both configurations have higher energy compared to non-Pro peptides. The corresponding energy difference between the *cis*- and *trans*-configurations is thus

smaller for an X-Pro peptide than for a non-prolyl peptide. Nevertheless, prolyl-isomerization can be a rate-limiting step during protein folding (Section 18.5.5). Most *cis*-peptides involving Pro residues are located at tight turns in loop regions, whereas non-prolyl *cis*-peptides are typically restricted to enzyme active sites or dimerization interfaces.

Until now, we have mostly focused on the conformation of the main chain. In Section 18.1.3 we will turn our focus onto the conformations sampled by the amino acid side chains.

18.1.3 Side-chain rotamers

Analysis of high-resolution protein crystal structures with well-defined electron density for the side chains has shown that the amino acid side chains adopt preferred conformations or *rotamers*. Rotamers are described by the torsion angles (Section 17.2.1.2), χ, of side chains. The χ_1 angle is the torsion angle around the Cα-Cβ bond, χ_2 corresponds to the rotation around the Cβ-Cγ bond, etc. The torsion angles of the rotatable single bonds along a side chain are clustered around values corresponding to staggered conformations (Section 17.2.1.2). In amino acids with short side chains, such as Ser, Cys, Thr, and Val, there are only three preferred conformations, each rotated 120° about the Cα-Cβ bond (Figure 18.5).

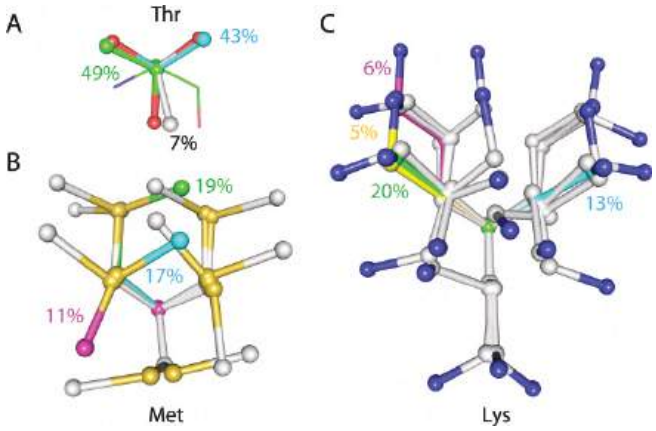


FIGURE 18.5: Preferred side-chain conformations in polypeptides. The view is along the Cβ-Cα bond onto the backbone of the polypeptide, and only the side-chain atoms are shown. The percentages for the most frequently observed rotamers in high-resolution crystal structures are given in the same color as the terminal carbon atom of the respective rotamer. A: Short side chains have three preferred rotamers separated by 120°. B: The 13 preferred rotamers for Met. C: Lys has 24 preferred rotamers.

The relative energies of these rotamers are determined by interactions with the main chain or neighboring side-chain atoms. In general, staggered conformations are associated with less steric hindrance, and are more stable than eclipsed conformations. Long side chains also show the basic three staggered conformations around the Cα-Cβ bond seen with the small side chains, but have additional degrees of freedom and are more flexible (Figure 18.5). Arg has 33 preferred rotamers, Lys and Met have 24 and 13 rotamers, respectively.

18.1.4 Post-translational modifications

The chemical space of proteins is increased by hundreds of post-translational modifications. Post-translational modifications can be covalent or non-covalent, permanent or transient, general for many proteins or specific to a single class of proteins. They occur spontaneously or are introduced by enzymes. Most post-translational modifications are introduced into unfolded proteins in the *endoplasmic reticulum* or the *Golgi apparatus*. Other modification reactions act on folded proteins, for example phosphorylation in the cytoplasm and ADP-ribosylation in the nucleus. Mass spectrometry (Section 29.1) has been invaluable for the identification of post-translational modifications, and the list of known post-translational modifications is continuously increasing. We will limit our discussion to a few frequently observed modifications and those that drastically change the biophysical properties of certain amino acids. Table 18.2 gives a short list of common amino acid modifications in proteins.

TABLE 18.2:
Summary of frequent post-translational modifications.

amino acid	modification	function / effect
N-terminus	deformylation of Met	removal of N-formyl group after translation
	cyclization of Gln	pyroglutamate in antibodies
	myristoylation	membrane anchor
Cys	oxidation to disulfide bonds	protein structure (see text), catalysis
	oxidation to sulfenic, sulfinic, sulfonic acids	sulfenic acids in catalysis, the others are irreversible oxidation products
	oxidation to formylglycine	active site residue of sulfatases
	binding of heme	prosthetic group for substrate binding
Asp	iso-aspartate	occasional shift of the peptide bond
Glu	γ-carboxylation	chelation of Ca^{2+} in blood coagulation factors
	ADP-ribosylation	
His	methylation, phosphorylation, cofactor attachment	
	conversion to diphtamide	
	ADP-ribosylation of diphtamide	in active site of translation elongation factor 2
Lys	methylation	mono-, di-, and trimethyl known
	hydroxylation and glycosylation	stabilizes and crosslinks collagen triple helix
	ubiquitination	signal for degradation by proteasome
	cofactor binding	binding of pyridoxal phosphate and retinal
	reversible histone acetylation	regulation of gene expression
Met	oxidation to sulfoxide	unspecific, but two diastereomers exist
Asn	glycosylation	protein folding and stability
	spontaneous deamidation	
Pro	hydroxylation plus glycosylation	hydrogen bonds in collagen
Gln	spontaneous deamidation	protein aging
Arg	methylation, phosphorylation, ADP-ribosylation, and hydrolysis to citrulline and ornithine	
Ser / Thr	glycosylation	mucus and extracellular matrix
	phosphorylation	signal transduction
Tyr	sulfation, phosphorylation	
	various phosphoesters with nucleotides	
C-terminus	methylation, amidation	removal of negative charge and preparation
	glycosyl-phosphatidylinositol	as membrane anchor
	farnesylation / geranyl-geranylation (Cys)	

The list is incomplete by necessity, only a few examples are listed per amino acid.

18.1.4.1 Glycosylation

Glycosylation is a common modification of proteins. The addition of carbohydrates introduces a large, highly hydrophilic moiety with extensive hydrogen bonding capabilities. Glycosylation occurs at the side chains of Ser, Thr, and Asn residues. The modification at Ser and Thr is termed *O-glycosylation*, at Asn residues it is referred to as *N-glycosylation*. Asn side chains are glycosylated when they are part of the N-X-[ST(C)] motif, where X is any amino acid except Pro (Figure 18.6). The transfer of single sugars or whole trees of carbohydrates is catalyzed enzymatically, and up to 500 different enzymes are involved in protein glycosylation. The glycans on a protein can easily double the protein mass, and substantially alter the biophysical properties of the protein. Glycosylation increases both the solubility and the stability of a protein, but does not alter its structure. It has been estimated that about half of all eukaryotic proteins are glycosylated. Most proteins that are located at the cell surface are glycosylated, creating a dense layer of glycan near the cell surface, the *glycocalyx*. Glycans are involved in intracellular protein transport, protein-protein interactions, and cell-cell contacts *via* their *extracellular matrix*.

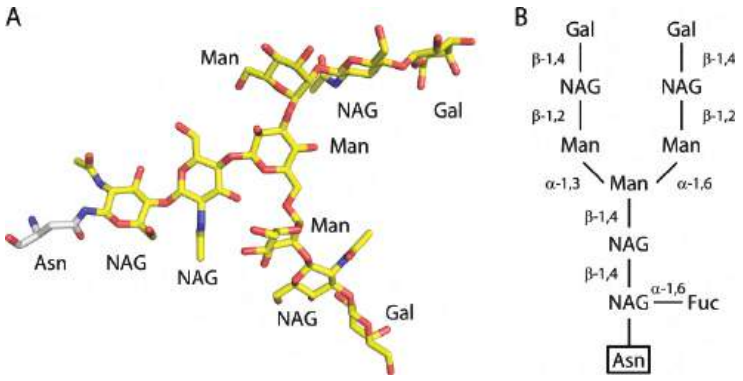


FIGURE 18.6: N-glycosylation. A: A complex glycan is connected to the side chain of Asn (gray). The formal connectivity of the monosaccharides is shown in (B). B: NAG is N-acetylglucosamine, Fuc is fucose (not shown in (A)), Man is mannose, and Gal is galactose.

Prokaryotes do not have enzymes of the glycosylation pathway, and prokaryotic proteins are not glycosylated. Recombinant production of proteins in *E. coli* that are post-translationally modified by glycosylation in their natural environment often leads to unstable and insoluble proteins.

18.1.4.2 Phosphorylation

Phosphorylation, the transfer of a phosphoryl group, is a widespread modification in both eukaryotes and prokaryotes. Phosphorylation introduces a small group with at least one extra negative charge that can engage in strong electrostatic interactions with other ions or dipoles, either in the same protein or in a binding partner. In eukaryotes, phosphorylation occurs at the side chains of Ser, Thr, or Tyr. In bacteria, Asp, His, and Lys may also be phosphorylated. Phosphoryl group transfer from ATP is catalyzed by enzymes termed *kinases*. The reverse reaction, the hydrolysis of the phosphoester, is catalyzed by *phosphatases* (Figure 18.7). The reversible modification of proteins with phosphate groups functions as a molecular switch: phosphorylation can either activate or inactivate a protein, and conversely dephosphorylation inactivates the protein and switches its activity off, or activates and switches it on. Phosphorylation is one of the most important and best studied post-translational modifications because it regulates a whole set of cellular processes including cell cycle, cell growth and metabolism, apoptosis, and signal transduction.

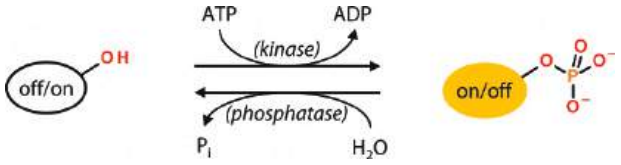


FIGURE 18.7: Protein phosphorylation. Phosphorylation of the hydroxyl groups of Ser, Thr, and Tyr can activate a previously inactive protein or *vice versa*. The phosphoryl donor is almost always ATP. In special cases Asp, His, and Lys are also phosphorylated.

Phosphorylated amino acid residues in proteins are recognized by specific protein domains (Section 18.3.4). For example, phosphotyrosine is recognized by the Src homology 2 (SH2; Figure 18.36) and the phosphotyrosine binding (PTB) domains. Phosphoserine and phosphothreonine are bound by WW (Figure 18.36) and fork head-associated (FHA) domains, respectively. By binding to phosphorylated residues in proteins, these domains mediate the assembly of protein-protein complexes only when their binding partner is phosphorylated, a strategy central to signal transduction across cell membranes.

Phosphorylation is also a means to modulate enzymatic activity. Glycogen phosphorylase becomes activated by phosphorylation and then releases glucose from its storage form glycogen. By

contrast, pyruvate dehydrogenase, an enzyme connecting glycolysis with the citric acid cycle, is less active when phosphorylated.

18.1.4.3 Hydroxylation

Hydroxylation is a modification that converts H into an OH group, introducing a polar functional group with hydrogen bond donor and hydrogen bond acceptor properties. Hydroxylation occurs at Pro and Lys residues (Figure 18.8) and is enzymatically catalyzed. It is a rather specific but important post-translational modification in collagen, where Pro and Lys residues are hydroxylated by the enzymes prolyl-4-hydroxylase, prolyl-3-hydroxylase, and lysyl-5-hydroxylase. Pro in XPG sequences in pro-collagen is hydroxylated at the 4-position to *trans*-4-hydroxyl-Pro, whereas Pro in GP sequences is modified to *trans*-3-hydroxyl-Pro. Lys is hydroxylated at the 5-position when part of an XKG sequence. In collagen, these hydroxylated residues engage in hydrogen bonds that stabilize the collagen triple helix (Figure 18.26). The collagen hydroxylases are Fe^{2+}-dependent and require ascorbate (vitamin C) to keep the iron in the reduced Fe^{2+} state. Scurvy, a disease caused by a lack of vitamin C, leads to less or no hydroxylation of collagen, and a reduced collagen stability in connective tissue because these hydrogen bonds cannot be formed.



FIGURE 18.8: Hydroxylated Lys and Pro. 5-hydroxylysine (5-HyL) can be further oxidized to introduce crosslinks between collagen chains. The methyl group in the hydroxyprolines (HyPro) is *trans* to the carbonyl group.

Formylglycine is a modified serine residue with an additional hydroxyl group at Cβ. The two hydroxyl groups can eliminate water to form a formyl group. The formyl group not only has different hydrogen bonding properties, but it also provides a highly reactive group for modifications (Box 18.1). Formylglycine is the catalytic residue of sulfatases, important enzymes in the hydrolysis of sulfate esters.

BOX 18.1: FORMYLGLYCINE.

Formylglycine (FGly, Figure 18.9) is the catalytic residue in an enzyme class called *sulfatases*. It is generated by oxidation of a cysteine (in eukaryotes) or serine (in prokaryotes) side chain with molecular O_2 while the sulfatase is still in its unfolded state. If sulfatases are not modified this way, they fold into their native structure but are inactive, which results in a variety of severe diseases. The oxidation reaction is catalyzed by the *FGly-generating enzyme* (FGE). Eukaryotic FGE recognizes the Cys in the sequence context C-[TSAC]-PSR, and oxidizes it in a Cu^{+}-dependent fashion to FGly. The active form of FGly is the hydrated aldehyde, a *geminal diol* that is coordinated to a Mg^{2+} ion in the active site of the sulfatases. FGly can therefore be viewed as a hydroxylated serine, and indeed it acts as the nucleophile for the hydrolysis of sulfate esters, as does a serine in phosphatases. Sulfatases and phosphatases both produce a covalent enzyme-substrate intermediate. In phosphatases the phosphoserine is simply hydrolyzed to regenerate the enzyme but in sulfatases a sulfoserine would be too stable for hydrolysis. Instead, sulfate is eliminated from the sulfo-FGly and then the geminal diol is regenerated by addition of water. A biotechnological application

(*Continued*)

of FGly is attaching the C-[TSAC]-PSR sequence at the N- or C-terminus of a protein, or within a loop region, which allows introduction of the unusual formyl group into the protein. The formyl group can then be coupled with an amine group of a fluorophore, a spin label, etc.

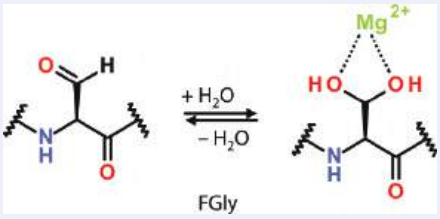


FIGURE 18.9: Formylglycine. In sulfatases, the formyl group is hydrated to a geminal diol that coordinates a Mg^{2+} ion.

18.1.4.4 Carboxylation

Carboxylation introduces a negatively charged group that often serves as an additional ligand for cations. An additional carboxyl group at the Cγ-atom of Glu (Figure 18.10) in proteins of the blood coagulation cascade introduces the ligands for binding of Ca^{2+}, which is essential for their function. Carboxylation of several glutamate residues occurs in the blood coagulation factors II (thrombin), VII, IX, and X. The carboxylase that catalyzes carboxylation of Glu to γ-carboxy Glu (Gla) is vitamin K-dependent. Warfarin and dicoumarol competitively inhibit the enzyme vitamin K epoxide reductase, thus depleting the vitamin K pool and ultimately rendering the carboxylase non-functional. These agents thereby inhibit blood clotting and act as anticoagulants. While warfarin was developed as a pesticide and medication against thrombosis, dicoumarol is a natural product that can lead to hemorrhage in livestock.

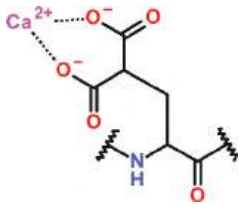


FIGURE 18.10: γ-carboxy glutamate (Gla). HCO_3^- is the carbon donor for the carboxylation. Normally, 1,3-dicarbonic acids are unstable but deprotonation and chelation by Ca^{2+} stabilizes Gla.

18.1.4.5 Disulfide bonds

A disulfide bond or *disulfide bridge* results from the oxidation of two Cys residues close in space. The cytoplasm is a reducing environment, and therefore cytoplasmic proteins usually do not contain disulfide bonds. In extracellular proteins, disulfide bonds are omnipresent. In eukaryotes, disulfide bonds are synthesized in the oxidizing environment of the endoplasmic reticulum, in bacteria in the periplasmic space. The covalent bond of disulfides can have a huge stabilizing effect on protein structure. Such stabilizing disulfide bonds are usually buried in the interior of the protein, inaccessible for reducing agents or disulfide-reducing enzymes. On the other hand, disulfides at the surface of proteins can be reduced to cysteines and are often involved in catalysis. Enzymes such as thioredoxin and glutaredoxin (Section 5.6.1) have two solvent-exposed Cys residues in their active sites that cycle between reduced (thiols) and oxidized (disulfide-bonded).

There are only a few preferred geometries for disulfide bridges, which are the left-handed spiral, the right-handed hook, and an extended structure found in immunoglobulins. These disulfide structures are described by the torsion angles χ_i when moving from the Cα of one Cys along the disulfide bond to the Cα of the other Cys. The preferred χ_1 angles of Cys (the torsion about the Cα-Cβ bond) are around −60° and 180°, with a strong preference for $\chi_1 = -60°$. The χ_3 torsion angle (around the Sγ-Sγ' bond) is in the range of ±90°, meaning that when viewed along the disulfide-bond, the Cβ-Sγ bonds are almost perpendicular to each other. If χ_3 is −90°, a left-handed spiral is formed by the atom sequence Cα-Cβ-Sγ-Sγ'-Cβ'-Cα' (Figure 18.11). The left-handed spiral is the predominant structure of disulfides in proteins. If χ_3 is +90°, the disulfide is right-handed and has the shape of a hook (Figure 18.11). The third kind of disulfide bond structure stabilizes the immunoglobulin fold (Figure 18.37) by crosslinking two β-sheets. This disulfide bridge is extended, with both the χ_1 and χ_2 angles close to 180°, such that the Cα atoms are maximally far apart from each other (up to about 700 pm).

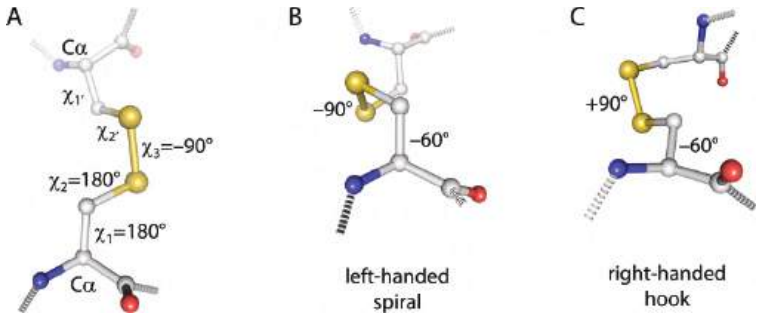


FIGURE 18.11: Disulfide bonds in proteins. The predominant disulfide bond structures found in proteins are the left-handed spiral, the right-handed hook, and an extended disulfide bond prevalent in immunoglobulins. The χ_i torsion definitions are noted for the extended disulfide bond in (A). The spiral (B) and the hook (C) both have χ_1 close to −60°.

18.1.4.6 Metal ion binding

Metal ions are bound to proteins by several side-chain ligands. As such, they constitute the second most prominent crosslink after disulfide bonds. Binding of metal ions adds a permanent charge, which engages in electrostatic interactions with other charges, and can bring negatively charged residues into close proximity. Metal ions in proteins serve three general functions: catalysis, regulation, and stabilization. Metalloenzymes incorporate one or more metal ions as part of their active sites, either alone or in complex with a co-factor. Examples are Mg^{2+} in sulfatases (Box 18.1), Zn^{2+} in alcohol dehydrogenase and zinc hydrolases, Fe^{2+}/Fe^{3+} in collagen hydroxylases (Section 18.1.4.3) and bound to heme in cytochromes, Ni^{2+} in urease, Cu^{2+} in ascorbate oxidase, and Mn^{2+} in arginase. Ca^{2+} is a regulatory metal ion in enzymes of the blood coagulation cascade (Section 18.1.4.4) and in calmodulin (Figure 18.36). Ca^{2+} bound to Gla residues in coagulation factors allows their binding to phospholipid membranes, which helps regulating the clotting process. Calmodulin is a sensor for Ca^{2+}, and changes its conformation upon Ca^{2+}-binding to activate or inactive target proteins involved in Ca^{2+}-signaling.

Proteins in general can be immensely stabilized by metal ion binding. The bond between a metal ion and its ligand is known as a *dative bond* or *coordinate bond* because the electron pair for the bond comes from the donor ligand. The electron-deficient metal ion is the acceptor for a lone pair from a donor atom located in a side chain, a peptide group, or a co-factor. Dative bonds are shorter than purely ionic interactions but they are longer (Table 18.3) and therefore not quite as strong as canonical covalent bonds (Table 17.1).

TABLE 18.3:
Metal ion distances to water and protein atoms.

Ion	H_2O	Asp/Glu-O	carbonyl-O	His-N	Cys-S
Na^+	241 ± 10	241 ± 10	238 ± 10	–	–
Mg^{2+}	207 ± 5	207 ± 10	226	–	–
K^+	281 ± 15	282 ± 15	274 ± 15	–	–
Ca^{2+}	239 ± 10	236 ± 10	236 ± 10	–	–
Mn^{2+}	219 ± 5	215 ± 5	219	221 ± 10	235
$Fe^{2+/3+}$	209 ± 10	204 ± 10	204	216 ± 15	230 ± 5
Co^{2+}	209 ± 10	205 ± 5	208	214 ± 10	225 ± 15
Cu^{2+}	213	199 ± 15	204	202 ± 10	215
Zn^{2+}	209 ± 5	199 ± 5	207	203 ± 5	231 ± 10

Bond lengths are given in pm ± standard deviations. Some metal ions have no standard deviation due to ill-defined statistics, variations in coordination number, or Jahn-Teller effects (*e.g.* unusually long axial bonds in Cu-complexes). The standard deviations are rounded values suggested for use in structure refinements (Section 24.4.6). Asp/Glu means a mono-dentate bond to a side-chain oxygen atom, carbonyl-O is a main-chain contact, and His-N and Cys-S are contacts to the respective side chains. Data compiled by Harding *et al.* using high-quality crystal structures of small-molecules and proteins (at least 1.25 Å resolution). Reproduced with permission. DOI 10.1080/0889311x.2010.485616.

Similar to hydrogen bonds (Section 17.2.2.3), dative bonds have partial covalent character. Estimating the degree of covalent bond character allows judging the stability of a solvent-exposed ionic interaction. The more ionic character an interaction has, the easier it dissociates in aqueous solutions. A strong covalent character of a dative bond is characteristic for permanent interactions, irrespective of solvent-exposure. The degree of covalent character depends on the physicochemical properties of both the ion and the ligand donating the electron pair. These properties include electronegativity, ionic radius, electron configuration, and polarizability. Metal ions found in biology follow the trend $K^+ < Na^+ < Ca^{2+} < Mg^{2+} < Mn^{2+} < Fe^{2+} < Co^{2+} \approx Ni^{2+} \approx Cu^{2+} < Zn^{2+}$ while the ligands can be ranked $CO_2^- < H_2O \approx$ OH-group of Ser/Thr ≈ amide oxygen atoms < His-N < Cys-S^-. Thus, interactions of alkaline metal ions with carboxylates and water are mostly ionic with little covalent character while copper and zinc ions interacting with the side chains of His and Cys have strong covalent character (see Section 19.5.4.1). The other combinations of ions and ligands lie in between these extremes.

The distances to Asn/Gln side-chain oxygen atoms are similar for all metal ions. Distances to Ser/Thr side chains are between those for water and Asp/Glu. Tyr-OH distances are > 10 pm shorter than those for Asp/Glu, and OH^- distances are 10–20 pm shorter than Asp/Glu. The distances actually found in proteins may deviate by up to three standard deviations (Table 17.1). Metal ions are bound with equilibrium dissociation constants in the range of pM (almost irreversible) to mM (loose interaction), leading to different degrees of stabilization. Na^+, Mg^{2+}, and K^+ are usually weakly bound to surface Glu/Asp side chains or the phosphate backbone of nucleic acids and readily exchange with other ions from the solvent. The most common strongly stabilizing metal ions in proteins are Ca^{2+} and Zn^{2+} (Figure 18.12). While Zn^{2+} usually binds irreversibly, Ca^{2+} can bind both reversibly and irreversibly, depending on whether it has regulatory or structural functions. Removal of structural metal ions greatly reduces protein stability and may lead to unfolding. For example, chelation of Ca^{2+} in subtilisin or calmodulin by EDTA or EGTA destabilizes these proteins by leaving a negatively charged cavity without compensating positive charges.

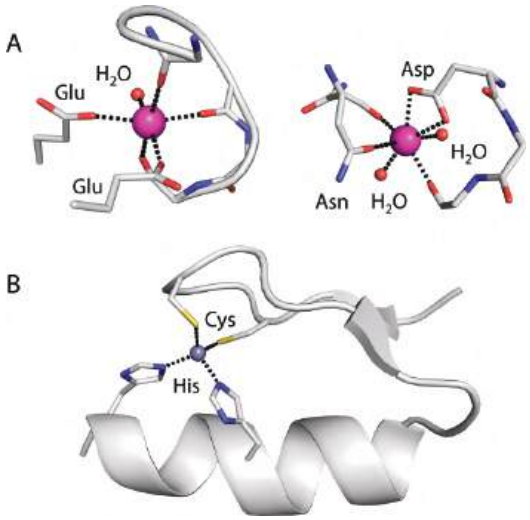


FIGURE 18.12: Metal binding in proteins. A: The two Ca^{2+} binding sites in the FGly-generating enzyme FGE (PDB-ID 1y1e). The coordinate bonds are shown as black dashed lines. Water molecules or hydroxide ions (these are difficult to distinguish) are shown as red spheres. Side chains binding to the Ca^{2+} (magenta spheres) are labeled. All other contacts are made by the carbonyl groups of main-chain peptide bonds. B: Tetrahedral Zn^{2+}-coordination by two Cys and two His side chains in the DNA-binding C_2H_2 zinc finger Zif268 (PDB-ID 1aay). The Zn^{2+} ion is shown as a gray sphere.

Depending on the type of ion, different coordination geometries are favored. The vast majority of metal ions are coordinated octahedrally with four ligands surrounding the metal ion at the center of a plane and the last two ligands perpendicular to the plane (left Ca^{2+} site in Figure 18.12). Softer, less electronegative ions, such as K^+ and Ca^{2+}, can accept more than six ligands (Figure 18.12) and the ligands may form a heavily distorted octahedron around the metal ion. Zn^{2+}, on the other hand, when bound to proteins strongly favors tetrahedral over octahedral coordination. If water is bound to Zn^{2+}, it has a reduced pK_A value and a higher propensity to form a hydroxide. Such Zn^{2+}-bound hydroxides are the nucleophiles in zinc hydrolases such as carboxypeptidase A and carbonic anhydrase.

18.1.5 Expansion of the genetic code

Post-translational modifications serve to expand the chemical scope of the natural amino acids, modulating their interactions, structures and stabilities, and catalytic activities. There are two further ways to diversify the twenty canonical natural amino acids. One is by natural incorporation of non-standard amino acids during translation *in vivo*, the other is *in vitro* translation using re-programmed mRNA.

During protein biosynthesis, a triplet of nucleobases in the mRNA, the *codon*, is recognized by an *anticodon* of a cognate transfer RNA (tRNA; Section 19.6.2). The tRNA carries an amino acid as an activated ester at its 3'-end, ready for incorporation into the growing polypeptide chain by the ribosome. Among the $4^3 = 64$ possible codons (Table 18.4) are three stop codons UAG (*amber*), UAA (*ochre*), and UGA (*opal*). Normally, they mark the end of the coding sequence, where translation stops and the polypeptide is released from the ribosome. Under special circumstances, stop codons can be decoded as the special amino acids selenocysteine and pyrrolysine. The tRNAs carrying these amino acids recognize the stop codon in the mRNA and suppress termination of translation, hence they are known as *suppressor tRNAs*.

TABLE 18.4:
Genetic code.

1st ↓	2nd base U		2nd base C		2nd base A		2nd base G		3rd ↓
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	**U**
	UUC		UCC		UAC		UGC		**C**
	UUA	Leu	UCA		UAA	*ochre*	UGA	*opal*	**A**
	UUG		UCG		UAG	*amber*	UGG	Trp	**G**
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	**U**
	CUC		CCC		CAC		CGC		**C**
	CUA		CCA		CAA	Gln	CGA		**A**
	CUG		CCG		CAG		CGG		**G**
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	**U**
	AUC		ACC		AAC		AGC		**C**
	AUA		ACA		AAA	Lys	AGA	Arg	**A**
	AUG	Met	ACG		AAG		AGG		**G**
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	**U**
	GUC		GCC		GAC		GGC		**C**
	GUA		GCA		GAA	Glu	GGA		**A**
	GUG		GCG		GAG		GGG		**G**

Met (marked in blue) and Trp are the only amino acids encoded by a single triplet. All other amino acids are encoded by several triplets or *synonymous codons*. Stop codons are colored in orange. They are named after gems for historic reasons.

The incorporation of selenocysteine (SeCys) requires a specific recognition element in the mRNA, a stem-loop structure close to an UGA codon that mediates the interaction with the specialized suppressor tRNA, together with a dedicated translation factor. In some bacteria and archaea, certain UAG codons can be suppressed by a similar mechanism to SeCys and translated into pyrrolysine (Pyl). Being directly encoded in the natural mRNA transcript of a gene by stop codons, SeCys and Pyl have been termed the 21st and 22nd amino acids. Their abundance in proteins is low, however. The selenol group of SeCys has a low pK_A value of ca. 5.5. It is found in redox-active proteins that require a very negative redox potential for their function such as glutathione peroxidase, thioredoxin reductase, and formate dehydrogenases (see Section 5.6.1). Pyl is essentially a lysine with an additional pyrroline ring, which is the active-site residue of a class of methyltransferases in methanogenic archaebacteria (Figure 18.13).

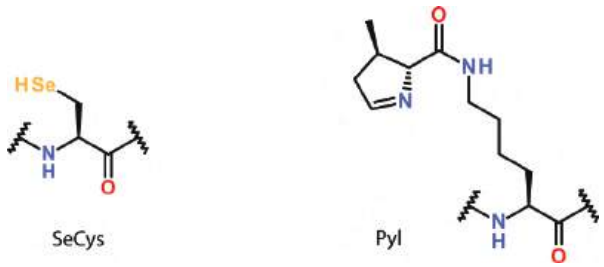


FIGURE 18.13: Selenocysteine (SeCys) and pyrrolysine (Pyl). SeCys differs from Cys by just a single atom, but with strong effects on the pK_A, polarizability, and redox potential of the side chain. Pyl incorporates a chiral pyrroline ring by an amide bond to the terminal NH_2-group of lysine.

The suppression of stop codons has been used with great success to introduce novel chemical functionalities into proteins. The re-programming is possible both *in vitro* and *in vivo*. In *in vitro* applications, the tRNA recognizing the respective stop codon is aminoacylated with a chemically altered, non-natural amino acid. Aminoacylation of tRNAs can be achieved chemically or enzymatically, using modified aminoacyl-tRNA synthetases with altered amino acid specificity.

Nowadays, evolved RNA molecules known as *flexizymes* (Box 19.3) catalyze tRNA acylation with a wide variety of non-natural amino acids with high efficiency. *In vitro* translation of the re-programmed mRNA in the presence of the aminoacylated tRNA then leads to incorporation of the amino acid at each stop codon. *In vivo*, the suppression of stop codons is possible with pairs of tRNAs recognizing the stop codon and aminoacyl-tRNA transferases that catalyze charging of the tRNA with the non-natural amino acid. In the last decades, a large set of such pairs has been generated for the *in vivo* incorporation of non-natural amino acid in bacteria, yeast, and human cell lines. The term *expanded genetic code* refers to the set of chemically altered, non-natural amino acids.

The degeneracy of the genetic code (Table 18.4), in which most amino acids are represented by 2–6 synonymous codons, offers the possibility to re-program mRNAs and incorporate non-natural amino acids at other codons as well. Using non-natural amino acids, protein function can be studied with homologs of natural amino acids or entirely different chemical moieties. Cys, Gln, and Met have been elongated by adding CH_2 groups. By introducing reactive groups such as ketones, boronic acids, or azides, proteins can be modified with NMR-sensitive isotopes, fluorophores, IR-sensitive probes, and spin labels (Figure 18.14). The chemical properties of these novel amino acids (keto, azide, boronic acids) are very different (*orthogonal*) from those of the canonical amino acid side chains (hydroxyl, thiol, amino groups) and allows labeling of proteins by two or more chemical groups, for instance in FRET studies (Section 25.7.6). Whole libraries containing hundreds of non-natural amino acids have been assembled to synthesize peptides with much more chemical variability than possible with the canonical amino acids only. Peptides containing non-natural amino acids can also be resistant to proteolysis. For example, substitution of 2-aminoisobutyric acids for an alanine renders the peptide hormone GLP-1 resistant to its natural inactivating protease DPP-4. A further change to the GLP-1 sequence (Lys28Arg) and a chemical modification of Lys20 have converted this hormone into the highly successful weight loss drug Semaglutide.

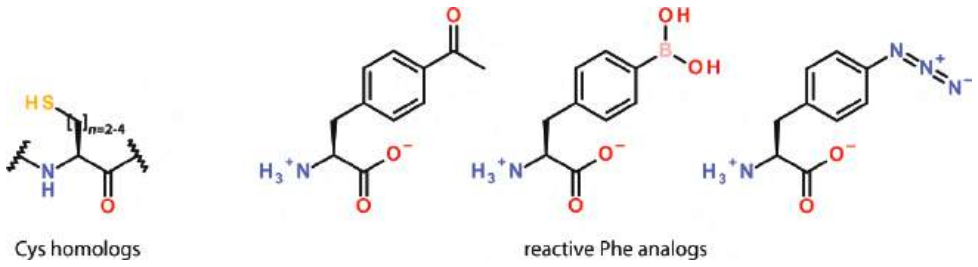


FIGURE 18.14: Homologs of Cys and Phe analogs. Homologs differ by the number of methylene (CH_2) groups. Analogs are modifications based on the same basic principle.

18.2 PROTEIN SEQUENCE

The sequence of a protein is also known as its primary structure. The individual properties of the amino acid side chains in a sequence sum to a set of overall properties of the resulting polypeptide. By analyzing the sequence, we can already learn a lot about the biophysical properties of the protein, including the propensity to fold, the charge and solubility at a certain pH, or its cellular localization. Protein sequences may be analyzed either in isolation, without any other source of information, or in conjunction with other sequences. By analyzing isolated sequences, global properties can be determined such as molecular mass, molar extinction coefficient, amino acid composition, frequency and distribution of amino acid residues, instability index, and isoelectric point (pI). By analyzing the sequence in small windows of a few amino acids, and moving this window along the sequence amino acid by amino acid (*sliding window analysis*), local parameters such as hydrophobicity, secondary structure propensity, or residue variability can be determined, providing information on solvent exposure, possible secondary structure, and flexibility. When several sequences are compared, similarities and differences become evident. Similar regions indicate conservation throughout evolution, often telling

about the function and/or stability of proteins. Different, non-conserved regions inform about areas that might be disordered or not involved in binding sites and function. Nevertheless, non-conserved regions contribute to the overall biophysical properties of a specific protein.

18.2.1 Analysis of individual sequences for global and local properties

The global parameters of an entire sequence may be calculated through web servers. If the calculated molecular mass from the sequence does not match the mass of a purified protein determined by mass spectrometry (Section 29.1), post-translational modifications (Section 18.1.4) may be deduced from the mass differences. Counting the number of absorbing amino acids allows calculation of the expected molar extinction coefficient of the unfolded polypeptide chain. Due to the dependence of absorption on the local environment, this value is similar, but not identical to the extinction coefficient of the folded protein (Section 21.2.4); both values typically differ by < 10%. More accurate values for the molar extinction coefficient of the folded protein require measurements using the Gill & von Hippel method (Section 21.2.7.1). The frequency of the twenty amino acid residues in a protein sequence can be compared to the mean distribution in a large set of sequences to gather information on its solubility and stability. If the frequency of hydrophobic amino acids is strongly increased from the average, this might point to a membrane protein (Section 18.4). For instance, bacteriorhodopsin from *Halobacterium halobium* has considerably more Leu/Gly and much less Glu/Lys than the average protein (Figure 18.15).

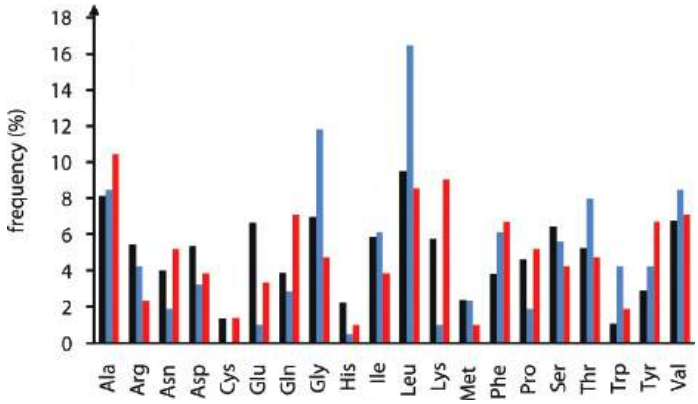


FIGURE 18.15: Residue frequency in a soluble and a membrane protein. The black bars are the average frequencies of the twenty amino acids in the UniProt database from 2013. The membrane protein *H. halobium* rhodopsin (blue) has more Gly and Leu residues than average, while the soluble *D. melanogaster* glutathione S-transferase (GST, red) more closely matches the average distribution.

The distribution of certain amino acids can also provide indications for regions that may be unstructured or intrinsically disordered (see Box 18.7). These regions are often composed of only a single or a few types of amino acids over stretches as long as 100 residues, termed *low complexity regions* (LCRs). An estimated 50% of all eukaryotic proteins is predicted to have disordered sequences of > 40 residues, while only 17% of prokaryotic proteins have such long disordered regions. Among LCRs that consist of a single repeating amino acid residue are Gln-repeats in the Parkinson's disease protein huntingtin (Table 18.13). Repeats of Ser, Gly, and Pro or mixtures of these residues can also signify unstructured regions. Modern protein modeling algorithms can identify disordered regions in non-repeating sequences that are otherwise difficult to spot by eye (Section 20.3).

The instability index is a measure for the *in vivo* resistance of a protein against proteolytic degradation. It is calculated based on the frequency of certain dipeptides in the sequence. Dipeptides have been assigned dipeptide instability weight values (DIWV), based on the comparison of the frequency of all 400 possible dipeptides in stable and unstable proteins. The tabulated values are used in calculation of the instability index of novel protein sequences. For a protein of length L, the

instability index is the sum of all DIWVs over all dipeptides $x_i x_{i+1}$ at position i along the sequence (eq. 18.1). Values < 40 predict a protein to be stable.

$$\text{instability index} = \frac{10}{L} \cdot \sum_{i=1}^{L-1} DIWV\left(x_i x_{i+1}\right) \qquad \text{eq. 18.1}$$

The identity of the amino acid at the N-terminus of a protein sequence is also a major determinant for *in vivo* protein stability and has predictive power on the estimated half-life of the protein in human, yeast, and *E. coli* cells. This relationship is called the *N-end rule*. If the N-terminal residue happens to be R, L, K, F, W, or Y, then the half-life of the protein in *E. coli* is very short, in the range of minutes. For the production of recombinant proteins, these residues should be avoided. By placing stabilizing residues at the N-terminus, the half-life of the protein in *E. coli* can potentially be increased to > 10 h. This effect is also taken into account when proteins are N-terminally tagged for affinity chromatography.

The number of charges of the protein can be calculated from the sequence as a function of pH. Strictly, this *in silico* titration is only valid for an unfolded protein where all ionizable side chains (Asp, Glu, His, Arg, Lys, termini) are fully solvent-exposed and their pK_A values (see Table 18.1) are not perturbed by the protein environment or the solvent. Since most of the charges of a protein are located at the surface, the pI of the folded protein is often within ±0.5 pH units of the calculated pI. The isoelectric point of the protein is the pH at which the overall charge is zero (Figure 18.16) and the solubility of the protein is minimal. This information can be used to keep proteins from aggregating during their purification and concentration by buffering at a pH very different from the pI.

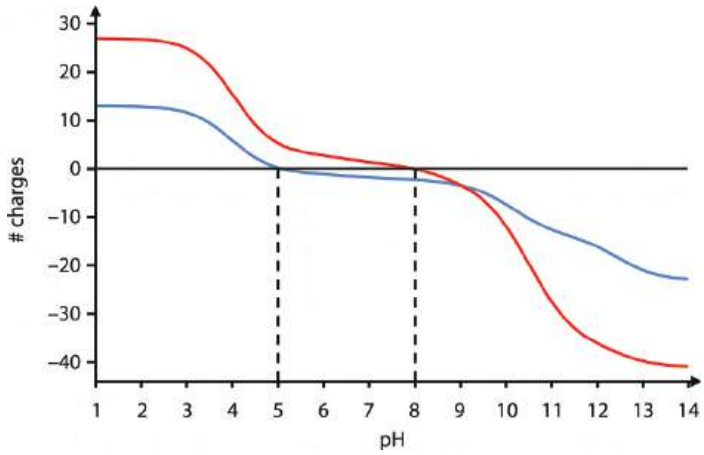


FIGURE 18.16: Titration of polypeptides. GST from *D. melanogaster* (red) and bacteriorhodopsin from *H. halobium* (blue) have quite different pI values (dashed lines at pH 8 for GST and pH 5 for bacteriorhodopsin), and different steepness of the titration curves around the pI. The membrane protein has a lower number of total charges than the soluble protein, especially in the physiological pH range.

The steepness of the titration around the pI can be a guide for how far away from the pI and in which direction the pH of the protein buffer should be. For example, *D. melanogaster* GST (Figure 18.16) has a pI of 8 and will accumulate fewer charges when shifting the pH to 6 (ca. −6 charges) than when shifting the pH to 10 (ca. +10 charges). In other words, increasing the pH of the buffer by one unit from the pI should lead to a higher solubility of the protein than decreasing it by one unit. Note, however, that counter ions are not taken into consideration when calculating the charges as a function of pH but if present in the buffer they may screen surface charges on the proteins. Experimentally, the best pH and most stabilizing buffer can be determined by using the calculated titration as a guide for a pH window and measuring the stability of the protein in a series of buffers in that pH window with *e.g.* the thermal shift assay (Box 18.8).

In addition to charges, exposed hydrophobic surfaces also determine the solubility of proteins. The Kyte-Doolittle and Sweet-Eisenberg hydropathicity schemes assign each amino acid a score expressing the relative hydrophobicity of its side chain (see Table 18.1). In the Kyte-Doolittle scale, Arg is the most polar side chain with a value of –4.5, and Ile is the most non-polar with a value of +4.5. Values for other schemes are different, but the relative values for the amino acids are very similar. The hydropathicity scores are averaged over a small sequence window in a sliding window analysis, and plotted as a function of sequence position (Figure 18.17). A continuous stretch of hydrophobic residues is a good indication that this part of the sequence is either buried in the hydrophobic core of the protein or embedded in a membrane.

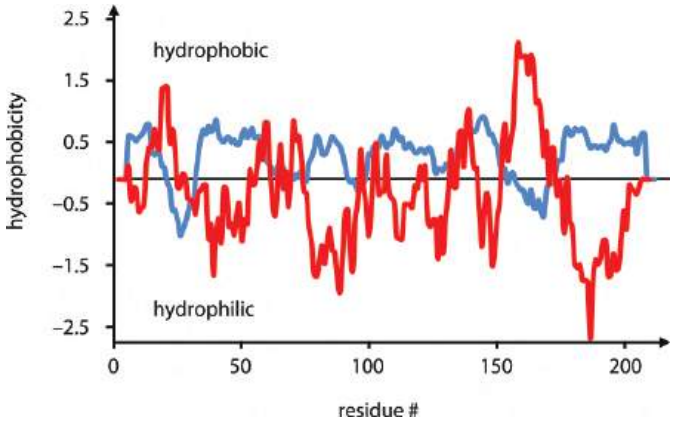


FIGURE 18.17: Kyte-Doolittle hydropathicity plot. The same examples as in Figure 18.15 are shown. Hydrophobicity values from Table 18.1 are assigned to each amino acid and averaged over a window of nine residues. Clearly, the membrane protein bacteriorhodopsin (blue) is overall more hydrophobic than the soluble GST (red). The hydrophilic parts of bacteriorhodopsin are the loops that extend away from the membrane on the cytosolic or extracellular side. Long hydrophobic stretches in GST belong to the hydrophobic core.

Finally, probable sites for ligand binding and post-translational modifications, including phosphorylation, glycosylation, and lipidation, can be identified by searching for characteristic short sequence patterns or *sequence motifs*. A few examples of motifs are given in the following: (1) cAMP-dependent kinase phosphorylates the Ser in the RRX_nS motif (n is usually 1, but 0 and 2 are also recognized). (2) The Tyr in [D/E]XXY motifs is phosphorylated by the anaplastic lymphoma kinase. (3) Ras-like GTP-binding proteins are farnesylated or geranylgeranylated for membrane attachment at the Cys of their C-terminal CaaX box, where the Cys (C) is followed by two aliphatic (aa) and an arbitrary amino acid (X). (4) The phosphate-binding loop or Walker A motif of canonic sequence GXXXXGK(T/S) is frequently found in nucleotide binding proteins. It forms a tight loop that orients its NH-groups toward the phosphates of nucleotides for hydrogen bonding (see Figure 18.42). The Lys side chain in the motif helps in neutralizing the phosphate charges, and the hydroxyl group of Thr or Ser binds to a Mg^{2+} ion, which in turn complexes the β- and γ-phosphates. Combinations of several motifs in the correct order within a sequence can help identify protein function. DEAD-box RNA helicases contain a Q-motif that binds the adenine base in ATP, followed by a phosphate-binding loop and a DEAD-sequence. Once this sequence of motifs is identified in a novel protein sequence, it is very likely an RNA helicase.

18.2.2 Sequence Alignment

Alignment of a novel protein sequence, the *query*, to several or all sequences in a database is a second important way to gain information from a sequence. Many protein sequence databases are categorized by organism, by organelles, or by diseases. Sequence searches over entire protein databases are done using online servers that run a variety of alignment algorithms. The first alignment algorithms

attempted to match all sequences over their entire length (Needleman-Wunsch) or to find local alignments by allowing gaps (Smith-Waterman). As the sequence databases became larger, less time-consuming algorithms using heuristic methods such as BLAST (Basic Local Alignment Search Tool) and FASTA were developed. All algorithms rely on scoring methods that assign positive scores to matching residues and negative scores to mismatching residues. Scoring of protein alignments is more complex and nowadays includes biophysical properties of the amino acid side chains such as polarity, hydrophobicity, charge, etc. (see Figure 18.1).

Many related proteins differ by insertions and deletions, differences in surface loop lengths, and insertions of whole domains. Such sequences cannot be aligned without introduction of gaps. If a gap has to be introduced into a sequence to optimize the alignment, a penalty is given by the scoring function. The value of the penalty may depend on where in the sequence the gap has to be introduced or how large it needs to be. Introducing a new gap is penalized heavier than extending an existing gap. Conserved sequences will form similar structures, while non-conserved sequences point to structural differences. Introducing a gap in a conserved region therefore is penalized more than in a non-conserved region of the sequence. While for low homology sequences different programs may result in different alignments, a consensus can often be reached when the results of several algorithms are compared. A very stringent way to align multiple sequences is to include the position of secondary structure elements, disulfide bonds, etc. in the alignment, provided that structures derived by NMR, X-ray crystallography, or cryo-EM for some of the proteins are available. Such *structure-based alignments* are more reliable than alignments that are based on sequences only (Figure 18.18).

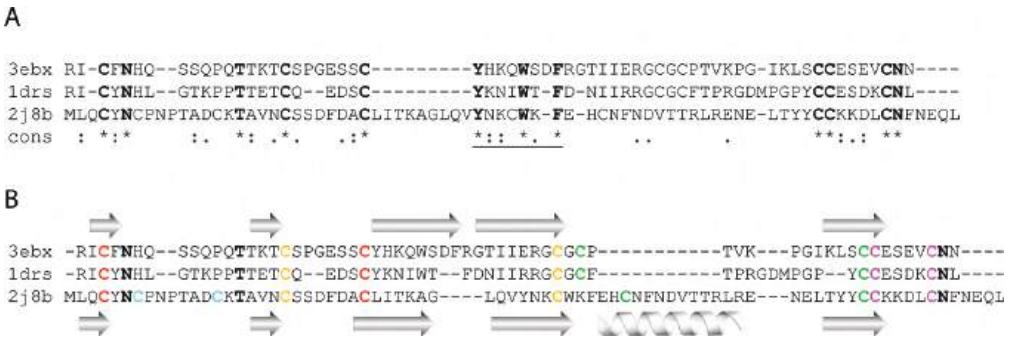


FIGURE 18.18: Multiple sequence alignment including structural data. Three proteins that belong to the snake toxin-like superfamily were aligned either just using their sequences (A) or including structural information (B). PDB-ID 3ebx is sea snake erabutoxin, PDB-ID 1drs is mamba dendroaspin, and PDB-ID 2j8b is an extracellular domain of the human cell surface receptor CD59. A: The snake toxins have quite similar sequences. The CD59 sequence has an insertion compared to the toxins that is placed by the alignment algorithm at a position to maximize the number of identical (*), very similar (:), and similar (.) residues. A particularly homologous region is underlined. B: All structures belong to the disulfide-rich fold with nearly all-β geometry. If this structural information is used, the alignment shifts significantly. The underlined homologous region from the top panel is gone, the insertion has shifted, and the pairs of Cys (colored) that form disulfide bonds can be assigned, showing that seven Cys positions, not six as in the top panel, are conserved. The secondary structure elements of the snake toxins and CD59, shown on the top and bottom, respectively, coincide very well. If shifts are present, they are only by one or two residues.

In general, multiple alignments of more than three sequences are more informative than pairwise alignments because evolutionary conserved residues can be identified from them. Sequence-based structure prediction algorithms (see Section 20.3.2) make extensive use of such *multiple sequence alignments* (MSAs), which sometimes contain thousands of sequences. The identification of conserved residues from MSAs helps in establishing phylogenetic relationships between proteins. Pairwise differences between all sequences are recorded in a matrix of *evolutionary distances*. The differences are scored by sophisticated methods that compare physico-chemical properties of the

side chains and take into account insertions and deletions, preferably mapped to crystal structures, if available. The results of the distance matrix are displayed in the form of a dendrogram that shows the probable diversification of the sequences during evolution. The dendrogram may or may not originate from a common ancestor. If no ancestor is identified the dendrogram is represented as a *rootless tree* (Figure 18.19).

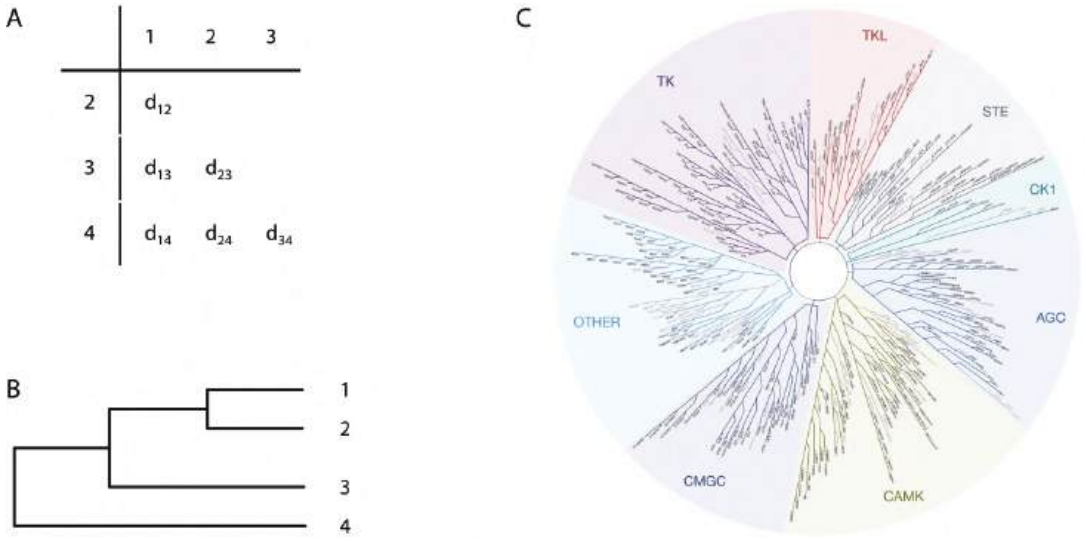


FIGURE 18.19: Phylogenetic relations between protein sequences. A: A distance matrix contains the similarities of all pairwise alignments. These distances can be visualized as a tree (B). C: More than 500 known kinases are grouped into eight families in a tree that has no common ancestor, a rootless tree. Image generated using TREEspot™ Software Tool and reprinted with permission from KINOMEscan®, a division of DiscoveRx Corporation, © DISCOVERX CORPORATION 2015.

In summary, the sequence contains a wealth of information on the overall biophysical properties, structure, function, stability, and evolutionary relationships of a protein. We will see next how the amino acid sequence dictates structure formation in proteins.

18.3 PROTEIN STRUCTURE

The primary structure of proteins is the sequence of amino acids connected by peptide bonds. This sequence contains all the information that is necessary to produce the final fold of the protein. Protein folding leads to the formation of different levels of structures (Figure 18.20): an unfolded polypeptide first adopts small, local structures, termed *secondary structure* elements, such as α-helix, β-strands, loops, and turns.

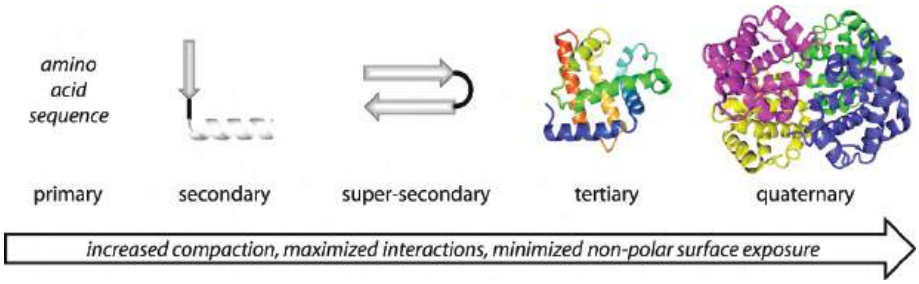


FIGURE 18.20: Different levels of protein structure. The primary structure is simply the amino acid sequence. Secondary structure elements are α-helix and β-strand, which are connected by loop regions. These elements, associated with characteristic backbone torsion angles, can form super-secondary structures or compact three-dimensional motifs. Secondary and super-secondary structures then form the compact tertiary structure, a globular protein domain. Association of several polypeptide chains into a complex is termed quaternary structure.

These secondary structure elements continue to fold into *super-secondary structures*, also called structural *motifs*. The next level of organization is the *tertiary structure*, in which secondary and super-secondary structures assemble and bring residues that are distant in sequence into close proximity to form *tertiary interactions*. Folded proteins are basically a string of secondary structure elements connected by loops and turns that pack tightly against each other to establish a three-dimensional structure. This structure can be a globular domain or an extended fiber. Proteins that further assemble into oligomers adopt a *quaternary structure*.

Secondary, super-secondary, tertiary, and quaternary structure formation follow two underlying principles: (1) maximizing the number of attractive interactions between atoms in the protein, and (2) minimizing exposure of hydrophobic protein surface to water. Both processes lead to the compaction of the molecule. Exposure of hydrophobic amino acid side chains to water is energetically unfavorable, and burying the hydrophobic side chains in the interior, away from the solvent, is a strong thermodynamic driving force for the folding of proteins into a compact structure. This phenomenon is known as the hydrophobic effect (Section 18.3.6.3). Interactions between atoms are established locally during the formation of secondary and super-secondary structures, and more globally during tertiary structure formation. Secondary structure elements have only a few attractive interactions and are not compact enough to bury much of the hydrophobic side chains. As a consequence, they have very limited stability in isolation, but become stabilized when they engage in tertiary and, optionally, quaternary interactions. Maximizing the number of attractive interactions requires changes in the main-chain and side-chain conformation, *i.e.* of torsion angles of chemical bonds. This change in conformation enables opposite charges to neutralize each other, dipoles to align to form hydrogen bonds, and non-polar surfaces to pack against each other for van der Waals interactions. We have seen in Section 17.1.2 that not all combinations of torsion angles or conformations are energetically equivalent. This is also true for amino acids and proteins: side chains have preferred rotamers, the most stable conformations with lowest energy (Section 18.1.3). The main chain also has preferred torsion angles, which give rise to specific secondary structure elements.

We have introduced the torsion angles ϕ and ψ that describe the conformation of the protein main chain in Section 18.1.2. A *Ramachandran plot* represents an energy surface as a function of ϕ and ψ (Figure 18.21). The energetically favored ϕ/ψ combinations are characteristic of the major secondary structure elements α-helix, β-strand, and turns. The α-helix corresponds to the energy minimum of $\phi = -60\pm10°$ and $\psi = -40\pm10°$, whereas the β-strand conformation has a large and shallow minimum around $\phi = -120\pm20°$ and $\psi = 125\pm15°$.

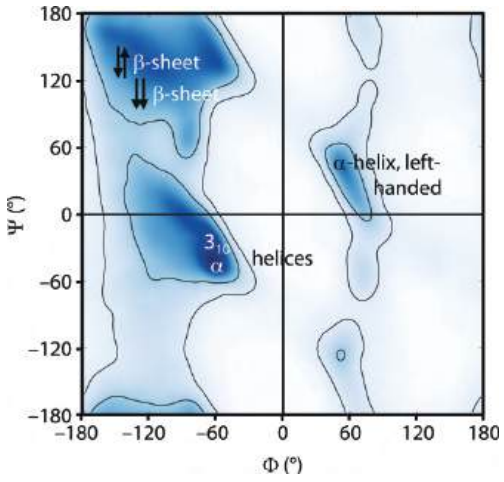


FIGURE 18.21: Ramachandran plot for non-Gly residues with secondary structure elements labeled. Darker areas are ϕ/ψ combinations of lowest energy while white areas indicate high energy and are therefore "forbidden". A left-handed α-helix is rare in natural peptides and proteins, but sometimes Gly adopts this backbone conformation. The figure was created with the programs Phenix and Molprobity.

Amino acids with different side chains prefer different ϕ/ψ combinations and thus have different secondary structure propensities (Table 18.5). For instance, the energetically favorable torsion angles of Ala are in the range of those compatible with α-helices. As a consequence, Ala is encountered much more frequently in α-helices than in any other secondary structure type. Pro and Gly are most frequently encountered in turns, and the small hydrophobic side chains Val and Ile are frequently found in β-sheets. Other residues, such as His, do not have a strong propensity for any particular secondary structure. The propensities of different amino acids for certain structures are used in secondary structure prediction algorithms (see Box 18.2).

TABLE 18.5:
Amino acid propensities for secondary structure.

amino acid	α-helix	β-strand	turn
E	**1.59**	0.52	1.01
A	**1.41**	0.72	0.82
L	**1.34**	1.22	0.57
M	**1.30**	1.14	0.52
Q	**1.27**	0.98	0.84
K	**1.23**	0.69	1.07
R	**1.21**	0.84	0.90
H	**1.05**	0.80	0.81
V	0.98	**1.87**	0.41
I	1.09	**1.67**	0.47
Y	0.74	**1.45**	0.76
C	0.66	**1.40**	0.54
W	1.02	**1.35**	0.65
F	1.16	**1.33**	0.59
T	0.76	**1.17**	0.90
G	0.43	0.58	**1.77**
N	0.76	0.48	**1.34**
P	0.34	0.31	**1.32**
D	0.99	0.39	**1.24**
S	0.57	0.96	**1.22**

Reproduced from Williams *et al.* (1987). BBA 916(2): 200–204, with permission. Bold numbers represent preferred secondary structure elements for these residues. A value of unity means no pronounced propensity for this secondary structure.

In Sections 18.3.1–18.3.3, we will have a closer look at the structural properties of secondary structure elements and the interactions stabilizing them.

18.3.1 Helical secondary structure elements

Helices are described by their radius R, the number of residues per turn r, and the *pitch* P, which is the height of a complete turn. The coordinates for the n^{th} residue of any helix may be calculated from these parameters according to eq. 18.2,

$$x_n = R \cdot \cos\frac{2\pi n}{r}$$
$$y_n = R \cdot \sin\frac{2\pi n}{r} \qquad \text{eq. 18.2}$$
$$z_n = \frac{P}{r} \cdot n$$

The ratio of the pitch and the number of amino acids per turn is also termed the *rise*. Using eq. 18.2, protein (and nucleic acid) helices of ideal geometry can be calculated for modeling into electron and potential density maps. In protein structures, the main types of helices are the α-helix, the 3_{10} helix, and the poly-Pro and collagen helices.

18.3.1.1 α-helix

The α-helix is a very compact secondary structure element (Figure 18.22) where, arithmetically, 3.6 residues form one turn of a right-handed helix with a pitch of 540 pm. Thus, each residue is rotated by 100° with respect to the next one. The right-handedness of the helix is a consequence of the Cα chirality of the amino acids. α-helices are stabilized by a regular pattern of hydrogen bonds between main-chain atoms: the carbonyl group of residue *i* hydrogen-bonds to the NH-group of residue *i*+4. As a consequence of these hydrogen bonds, the planes of the peptide bonds are aligned parallel to the α-helix axis. The individual dipoles of each peptide bond therefore sum to a macroscopic *helix dipole* with about half an electronic charge at either end. This macro-dipole can undergo strong polar interactions. Helix dipoles often stabilize anions bound near the helix N-terminus in proteins (see Figure 18.42 in Section 18.3.7). In α-helices, residues that are not adjacent in sequence become close in space, which can change surface properties and possible interaction patterns. Because of the *i*,*i*+4 hydrogen bonds, the side chains of residues *i* and *i*+4 lie on the same side of the helix. If side chains of the pattern *i*,*i*+4,*i*+7/8, etc. are hydrophilic, they form a hydrophilic patch along one side of the helix. Likewise, if these side-chains are hydrophobic, they generate a hydrophobic side. Such helices are termed *amphiphilic* (equivalent to *amphipathic*), meaning that they have hydrophilic residues on one side and hydrophobic residues on the other. The hydrophilic side of amphiphilic helices is typically exposed to water on the surface of proteins. The hydrophobic part of amphiphilic helices is oriented toward the interior of proteins or can interact with one leaflet of a membrane.

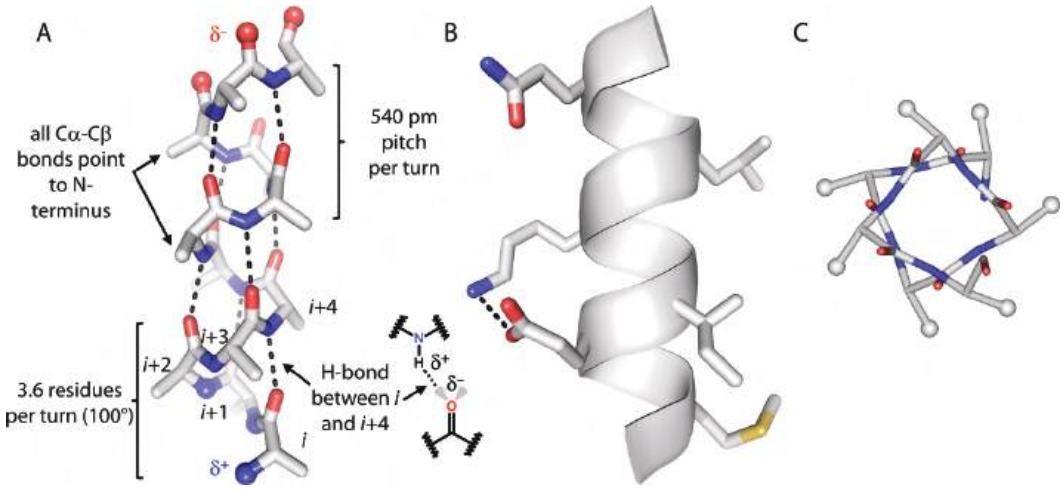


FIGURE 18.22: Geometry and properties of an α-helix. A: Stick representation of a poly-Ala sequence. The dashed lines represent the main chain hydrogen bonds between residues *i*,*i*+4. Three N-terminal nitrogen atoms and three C-terminal oxygen atoms that do not have a hydrogen-bonding partner are shown as spheres. The helix dipole is indicated by the δ^+ and δ^- labels. B: Ribbon representation of an amphipathic α-helix with polar side chains on the left-hand side and non-polar side chains on the right-hand side. A charged hydrogen bond between a Glu-Lys pair is shown as a dashed line. C: The top view of the α-helix shows staggered side chains.

Because of the *i*,*i*+4 hydrogen bond pattern, the first three NH-groups and the last three carbonyl groups of α-helices have no hydrogen-bonding partners (Figure 18.22), which is energetically unfavorable (Section 17.2.2.3). Free hydrogen-bonding valences at the termini of α-helices are often *capped* by amino acids just outside the actual helix. These residues at the position half-in and half-out of an α-helix are termed the N-cap and C-cap, respectively (Figure 18.23).

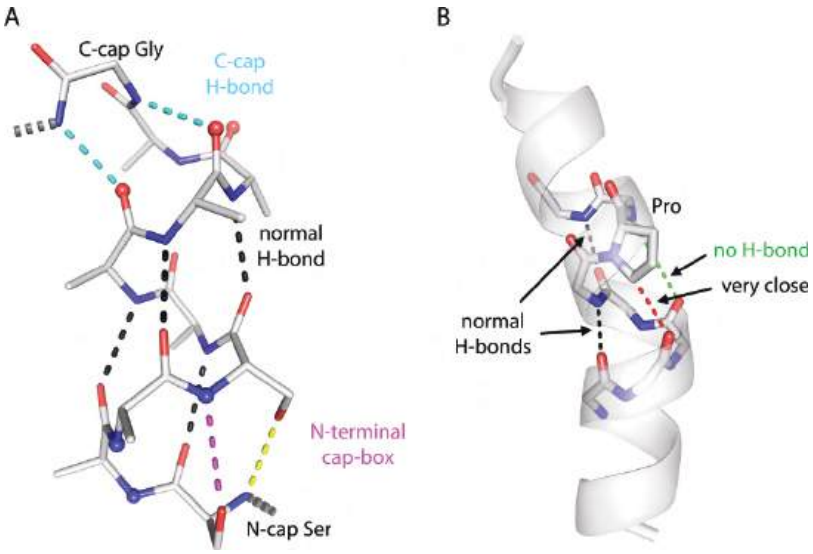


FIGURE 18.23: Helix capping and kinking. A: The Gly at the C-terminus of the α-helix alters the direction of the polypeptide chain. The chain would normally follow the right-handed spiral but escapes to the left. The peptide groups on either side of the Gly are tilted toward the helix axis to form two hydrogen bonds in inverted sequence order (cyan dashes). The N-cap shown here is a Ser whose side chain caps the NH-group of the third residue of the α-helix (magenta dashes). The side chain of this third residue accepts a hydrogen bond from the N-cap residue (yellow dashes), completing the cap-box. PDB-ID 1lmb. B: Pro can kink helices. Pro in the middle of a long α-helix leads to a kink and change of direction of the polypeptide chain. Because of the covalent bond between the Cδ atom of Pro and the main-chain nitrogen, no hydrogen bond is possible (green dashes). The Cδ atom has an unfavorably close contact with the carbonyl group of the *i*–4 residue (red dashes), which induces a kink in the helix. The α-helical hydrogen-bonding pattern is reached again at the *i*+2 residue.

Of all residues proline has the least propensity to occur in an α-helix (see Table 18.5). Although Pro in polypeptides can in principle adopt main-chain torsion angles compatible with α-helices, the Pro side chain lacks an amide proton (Figure 18.1) and cannot engage in an *i*,*i*+4 hydrogen bond. In addition, it would clash with the carbonyl oxygen atom of the preceding residue. Thus, helices often terminate at a Pro, and Pro is therefore colloquially termed a *helix breaker*. However, in the middle of longer helices, Pro residues do not strongly destabilize the entire helix but introduce a kink (Figure 18.23). Such kinks can be of structural and functional importance and are found in many G-protein coupled receptors (Section 18.4.4), which are membrane proteins of seven transmembrane α-helices where the often kinked helices pack together to form a ligand binding site.

Short isolated α-helices are not very stable in aqueous solvent; it is the packing of α-helices against each other that stabilizes them. The interaction between two helices can be described as a "knobs-in-holes" mechanism, where the side chains interdigitate. This interdigitation imposes restraints on the relative arrangement of the interacting helices: if two helices are inclined by 20° relative to each other and are overwound slightly to a helical repeat of 3.5 residues, every 7th residue of one helix will fit into a "hole" on the surface of the other. The helices wind around each other in a left-handed superhelix (Figure 18.24). This *coiled-coil* structure can often be predicted from a sequence that contains *heptad repeats*, *i.e.* a similar sequence every seven residues (two helical turns) that end up on the same side of the helix. The side chains that form the knobs on one helix and the walls of the holes on the other helix are mostly hydrophobic, so a coiled-coil is an efficient way to bury hydrophobic surfaces. Leu predominates, but other aliphatic side chains, and the larger aromatic side chains are sometimes found in the interface of coiled-coils. Leu-rich regions forming coiled-coils are also known as a *Leu-zipper*, a classical protein dimerization module that is, among others, found in dimeric DNA-binding proteins (Section 19.5.4.1).

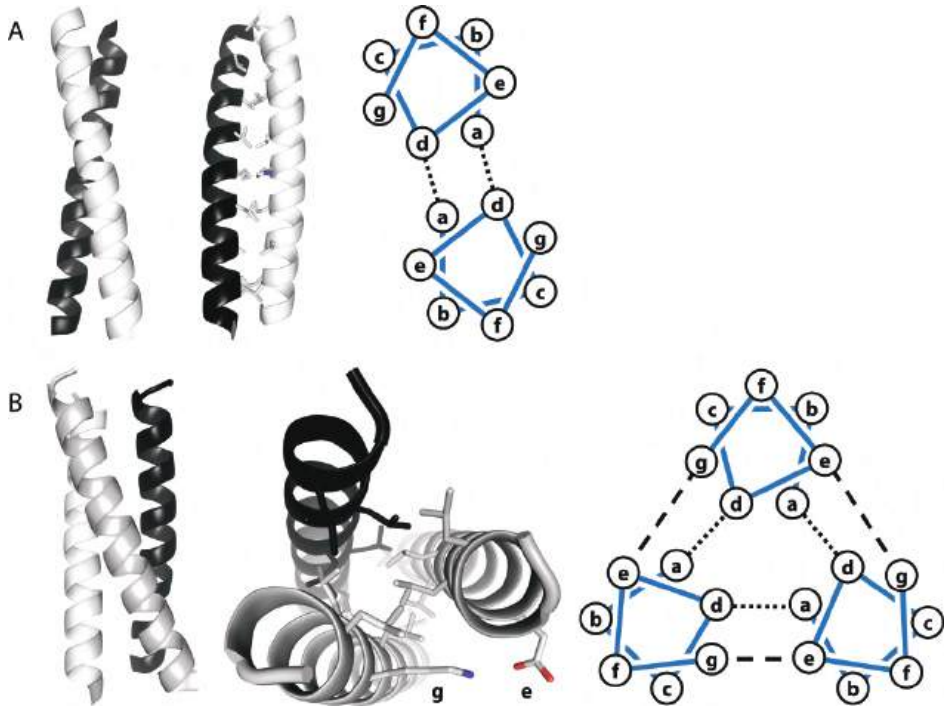


FIGURE 18.24: Coiled-coils of α-helices. A: The Leu-zipper is a classical parallel two-helix coiled-coil where the slightly distorted α-helices with a helical repeat of 3.5 residues assemble into a left-handed superhelix. The interface is hydrophobic. The schematic drawing on the right is the helix wheel representation of the heptad repeat viewed from the C-terminus. It shows how side chains from residues **a** (most frequently Leu, Ile, or Ala) and **d** (most frequently Leu or Ala) pack against each other. B: A three-membered parallel coiled-coil.

The helices in a coiled-coil can be parallel or antiparallel, with two, three, and four helices forming the bundle. Larger coiled-coils with more helices can be viewed as assemblies of the smaller 2-, 3-, and 4-membered coiled-coils. A schematic visualization of coiled-coils is the *helix wheel*. A single helix wheel shows the heptad repeats, and several helix wheels next to each other show the interacting residues in the coiled-coil.

In principle, the knob-and-hole pattern also interlocks for α-helices oriented at an angle of about 70°. This almost perpendicular helix-crossing is stabilized by fewer interactions, however, and is therefore less stable than the coiled-coil. However, it is often found as part of a larger structure where the two crossed helices are stabilized by other interactions.

18.3.1.2 3_{10}-, poly-Pro, and collagen helices

Apart from the α-helix, there are several more helical secondary structure elements that highlight the structural plasticity of the polypeptide chain: the 3_{10}-helix, the poly-Pro helix, and the collagen helix. The 3_{10}-helix follows an $i,i+3$ hydrogen bonding pattern along the main chain instead of the $i,i+4$ pattern in the α-helix. The 3_{10}-helix has a smaller radius than the α-helix, with tighter turns of the main chain, and is therefore more densely packed (Table 18.6). The dipoles of the amide groups are not aligned head-to-tail, leading to non-linear hydrogen bonds between residues i and $i+3$ that contribute little stabilization energy. Due to the fact that the 3_{10}-helix has three residues per turn, the side chains i and $i+3$ are eclipsed when viewed along the helix axis (Figure 18.25), leading to conformational strain. As a result, the 3_{10}-helix is less stable than the α-helix. The longest 3_{10}-helices in natural proteins have just two turns. They are often found at the end of α-helices where they contribute capping residues (Figure 18.23). The 3_{10}-helix got its name from the fact that the ring closed off by the hydrogen bond (including the hydrogen) involves ten atoms from three residues along the sequence (Figure 18.25). By this definition, the α-helix is technically a 3.6_{13}-helix, although this nomenclature is rarely used.

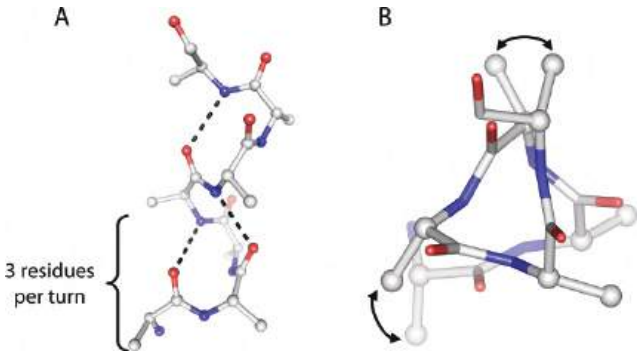


FIGURE 18.25: The 3_{10}-helix. A: The main chain is tightly wound. The *i,i*+3 hydrogen bonds (black dashes) are 300 pm long, a distance close to the optimum, but they have sub-optimal geometry. B: The side chains in the 3_{10}-helix emanate every 120° from the helix axis and are almost eclipsed (double arrows; compare to Figure 18.22). The tightly wound main chain and the close contacts of the side chains lead to conformational strain.

We have seen before (Figure 18.23) that Pro is a helix breaker. However, poly-Pro stretches can form helices that are not stabilized by regular hydrogen bonds. Instead, the driving force for their formation is the minimization of steric clashes of the conformationally restricted proline side chains. There are two types of poly-Pro helices. Type I has only *cis*-peptide bonds and forms in organic solvents, but has not been found in proteins. By contrast, type II has all peptide bonds in *trans* and is the predominant form in water. In contrast to all other types of helices, the poly-Pro-II helix is left-handed (Figure 18.26). It occurs in short linker regions connecting individual domains of multi-domain proteins. SH3-domains (Figure 18.32) and WW domains (Figure 18.36) specifically bind poly-Pro-II helices, which is important for the assembly of multi-protein complexes.

Very long poly-Pro-II helices are the basic structural elements in collagen, a ubiquitous structural protein that is part of tendons, ligaments, skin, and cartilage. About a dozen closely related variants of collagen exist, all of which have the basic sequence $(\text{Gly-Xaa-Yaa})_n$, where Xaa and Yaa are often Pro. Many of the Pro and also Lys at position Yaa are post-translationally hydroxylated to 3-hydroxy-Pro (HyPro) and 5-hydroxy-Lys (Section 18.1.4.3). Three poly-Pro-II helices associate into a parallel trimer (Figure 18.26) termed the *collagen helix*.

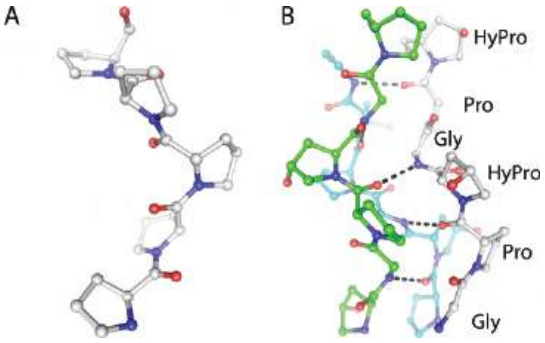


FIGURE 18.26: Helical secondary structure elements that involve Pro. A: Poly-Pro-II helix. The peptide bonds are *trans*, and the helix is left-handed. In a poly-Pro sequence no hydrogen bonds exist between residues along the helix because no hydrogen bond donor is present. B: Collagen triple helix made by three chains of poly-Pro-II geometry (PDB-ID 4dmt). Hydrogen bonds between the chains are made between Gly of one chain and residue Xaa (not Pro) of another chain in the Gly-Xaa-Yaa repeat.

In this arrangement, the individual helices periodically come very close. This is facilitated by a conserved glycine residue in every third position in the repetitive $(\text{Gly-Xaa-Yaa})_n$ sequence. The triple helix is stabilized by cross-strand hydrogen bonds from NH-groups of Gly with the CO-groups of Pro. Additional stabilization is provided by covalent crosslinking of the chains with hydroxylysine

and by hydroxyproline (Figure 18.8), although not by direct hydrogen bonds as the OH-groups are too far away from other chains.

Table 18.6 summarizes the geometric properties of the helices that we discussed in this Section.

TABLE 18.6:
Geometric properties of helical polypeptide conformations.

helix	α	3_{10}	poly-Pro-I	poly-Pro-II
handedness	right	right	right	left
# residues per turn	3.6	3.0	3.3	3
pitch (pm)	540	600	630	930
radius of helix (pm)	230	190	–	70*
preferred ϕ	–60	–75	–83	–75
preferred ψ	–45	0	158	150
hydrogen bond	*i*,*i*+4	*i*,*i*+3	–	–
# of atoms in ring	13	10	–	–

* The 70 pm radius of the poly-Pro-II helix refers to the collagen triple helix.

18.3.2 β-STRANDS AND THEIR SUPER-SECONDARY STRUCTURES (β-SHEETS)

The second important structural element of proteins is the β-strand, a maximally extended polypeptide (Figure 18.27). The extended conformation is a preferred structure; it corresponds to a broad energetic minimum in the Ramachandran plot at ϕ/ψ combinations of roughly –130°/130°. There are two regular patterns in a β-strand. First, subsequent side chains follow an up-down pattern, meaning every second side chain points to the same side. This arrangement determines the surface properties of the strand, which, similar as in amphiphilic helices, can be different for both sides. Second, the peptide groups have a back-and-forth orientation, with every second peptide group pointing into the same direction. This arrangement has consequences for the hydrogen bonding patterns possible between β-strands. Overall, the repeating unit of a β-strand is therefore a di-peptide with the Cα-atoms separated by a distance of ca. 680 pm.

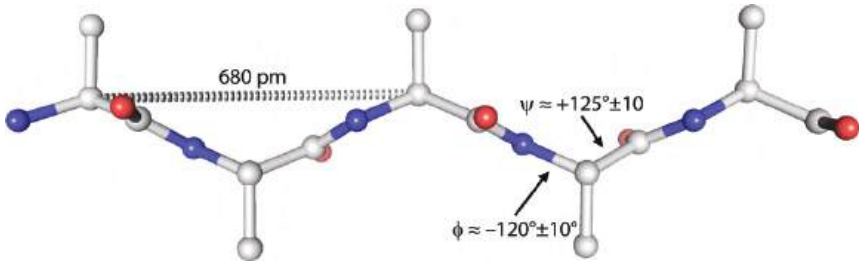


FIGURE 18.27: Ideal β-strand conformation. In a single ideal β-strand, the side chains alternatively point up and down. The main-chain torsion angles are roughly at ϕ/ψ combinations of –130°/130°. This corresponds to a shallow minimum in the upper left corner of the Ramachandran plot (Figure 18.21).

In contrast to α-helices where local *i*,*i*+4 interactions stabilize each turn, an isolated β-strand is an extended structure with minimized steric clashes, but no stabilizing interactions, and hence only marginal stability. However, β-strands can assemble into more stable structures, including β-*hairpins*, β-*sheets*, and β-*barrels*. Formation of these structures is mediated by hydrogen bonds between the peptide bonds of adjacent strands, maximizing inter-strand interactions and stabilizing the resulting structure.

Interactions between β-strands can be parallel, with the backbone of the interacting strands running in the same directions, or antiparallel, with opposite backbone directions. The smallest structure formed by the interaction of two β-strands connected by a short loop is the β-hairpin (Figure 18.28). Isolated hairpins can further assemble into super-secondary structures, including

β-*meander*, *Greek key*, and *jellyroll* (Figure 18.28). The β-meander topology is a building block in antiparallel assemblies of β-strands. The Greek key motif is an assembly of two β-hairpins into a four-stranded antiparallel arrangement found in immunoglobulin domains (Figure 18.37). The jellyroll is an extension of the Greek key motif with an additional β-hairpin attached to one side. It is prevalent in many spherical virus coat proteins.

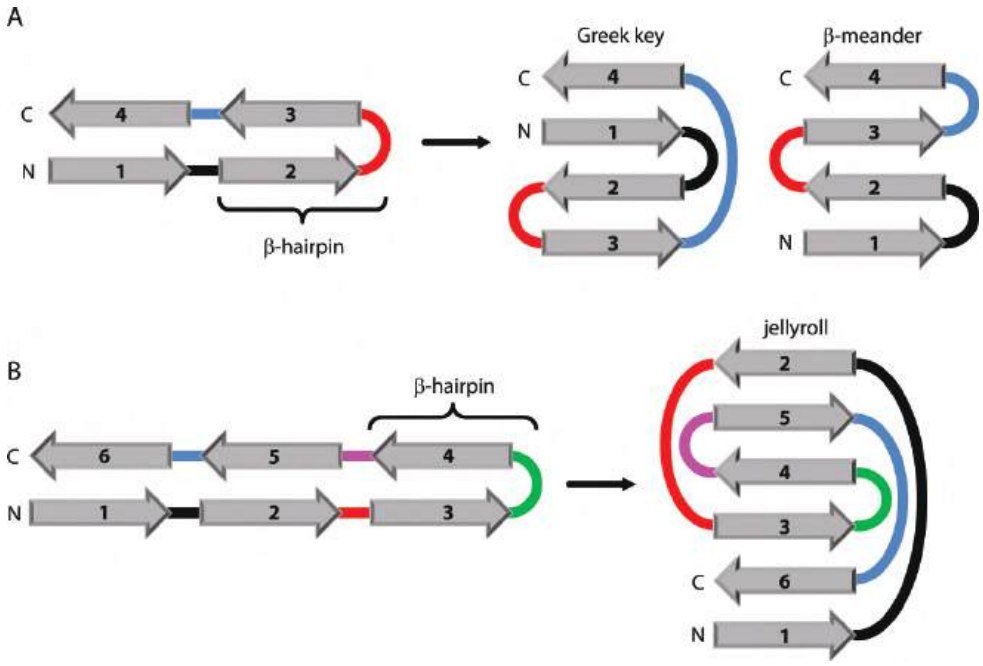


FIGURE 18.28: Super-secondary structures formed by β-strands running antiparallel. A: The β-hairpin forms from two consecutive β-strands that are connected by a short turn. Four β-strands can form the regular up-and-down arrangement of a β-meander, or a Greek key motif (named after Greek ornamental artwork). A total of 24 parallel and antiparallel arrangements with different connections can be drawn from four β-strands. The β-meander and Greek key motif are by far the most frequently observed β-motifs in protein structures. B: Six strands are often found to pack into the jellyroll motif: the three β-hairpins made up by strands 3/4, 2/5, and 1/6 are rolled up into a spiral.

With the Greek key motif, we have now encountered a first example of a β-*sheet*, a more extensive side-by-side interaction of several β-strands. β-sheets can be parallel, with all β-strands pointing into the same direction; antiparallel, with alternating directions of their β-strands; or *mixed*, with parallel and antiparallel strands (Figure 18.29). In these arrangements, the peptide bonds of the β-strands at the rim are not engaged in hydrogen bonds within the β-sheet, but can engage in other interactions.

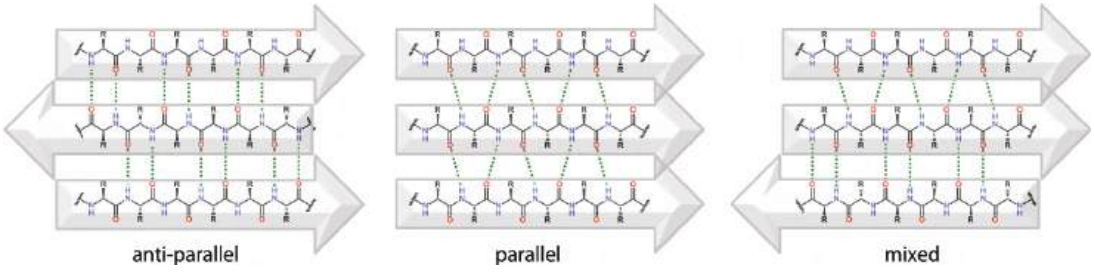


FIGURE 18.29: β-strands associate into antiparallel, parallel, and mixed β-sheets. The directions of the polypeptides from the N- to the C-terminus are shown as arrows. Hydrogen bonds (green dashes) are formed between the peptide groups of adjacent β-strands and have almost perfect geometry. Note that the hydrogen bonds in antiparallel β-sheets are parallel to each other and come in pairs while in parallel β-sheets they are neither parallel nor spaced evenly.

β-sheets are pleated: the Cα atoms of subsequent residues are slightly above and below the plane defined by the sheet. Together with the alternating upward- and downward-pointing side chains, "ripples" perpendicular to the direction of the polypeptide chain traverse the sheet, bringing residues in neighboring strands into close proximity. The side chains forming these ripples interact *via* inter-strand hydrogen bonds or van der Waals interactions (Figure 18.30). Mathematically, the β-strand can be described as a helix with two residues per turn and a translation of 320–340 pm per residue (Table 18.7). The entire β-sheet can be planar or twisted. β-sheets have a right-handed twist when viewed along the direction of the strands.

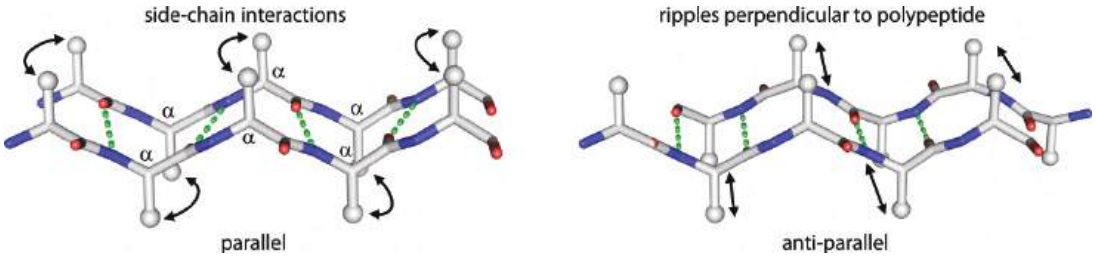


FIGURE 18.30: Parallel and antiparallel β-sheets. Hydrogen bonds are shown as green dashed lines. Hydrogen bonds in parallel β-sheets are angled with respect to each other while antiparallel β-sheets have parallel hydrogen bonds. Common to both β-sheets are side chain interactions between neighboring strands (double arrows) and the formation of ripples: the up-and-down path followed by the Cα atoms gives rise to ripples or pleats perpendicular to the polypeptide chains (straight double arrows).

TABLE 18.7: Geometric properties of β-strand conformations.

strand	⇅ β	⇈ β
# residues per turn	2	2
pitch (pm)	680	640
preferred ϕ	–140	–119
preferred ψ	135	113
hydrogen bond	long-range	

The shortest connections that are possible between the individual strands of a β-sheet depend on the nature of the sheet. In antiparallel sheets, the end of one strand and the start of the next strand in sequence are on the same side of the sheet, in close proximity. The connections therefore can be as short as two residues (a reverse turn, Section 18.3.3). For parallel β-sheets, the end of the first strand and the start of the second strand are on opposite ends of the sheet. As a consequence, a minimum number of residues is required to connect these strands. This can be a long loop (β-loop-β motif), or an α-helix (β-α-β motif). β-loop-β and β-α-β motifs are the building blocks of *Leucine-rich repeat* (see Figure 18.36) and *TIM-barrel* proteins (see Figure 18.37). The *crossovers* can be in two different configurations, left- and right-handed. The vast majority of crossovers are right-handed, a consequence of the inherent right-handed twist of β-sheets (Figure 18.31).

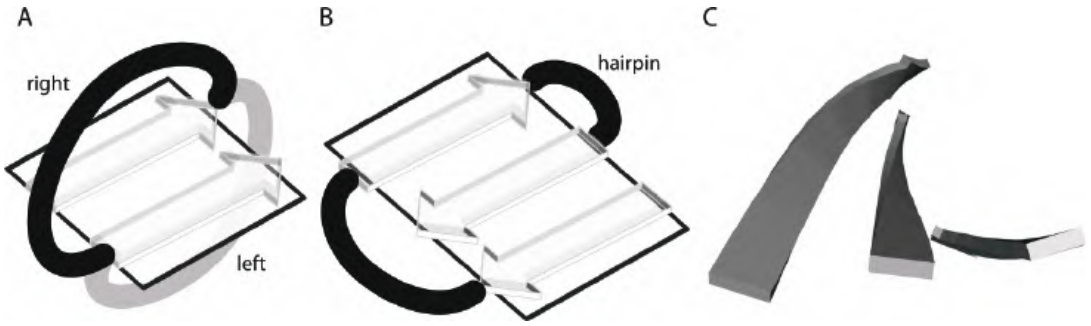


FIGURE 18.31: β-strand connections and bending of sheets. A: Crossovers connect β-strands at opposite edges of the sheet. The two strands of a parallel β-sheet and the crossover connecting them form a spiral with a defined handedness. The right-handed crossover is by far the dominant connection. The left-handed crossover (light gray) is very rarely observed. B: Hairpins connect adjoining strands on the same side of the β-sheet. C: Most β-sheets display a right-handed twist when viewed along the direction of the strands. Several more strands may finally close into a β-barrel.

Planar β-sheets or β-sheets with twists close to zero can grow infinitely. Small peptides are known to assemble into infinite sheets and form filaments in a number of diseases including Alzheimer's disease (see Table 18.13). Finite β-sheets expose hydrogen bond donors and acceptors at their first and last β-strand. These unsatisfied hydrogen bond donors and acceptors are destabilizing if the β-sheet is buried in the hydrophobic interior of a protein. If the first and last strand of a β-sheet interact with the hydrogen bond pattern characteristic of β-sheets, a closed, quasi-infinite β-sheet is formed, termed β-*barrel*. Formation is possible for β-sheets of at least five strands (Figure 18.32). The strands in a β-barrel may run parallel or anti-parallel to each other, just like in any β-sheet. If β-barrels have uneven strand number, the first and last strands are parallel by necessity. The strands in a barrel are inclined relative to the main barrel axis. The fewer strands a barrel has, the more inclined they tend to be (Figure 18.32). Some barrels, such as SH3 domains and GFP, are stable on their own. Such β-barrels are either entirely antiparallel or are mixed parallel and antiparallel. By contrast, barrels with exclusively parallel β-strands necessarily have long crossovers on their outside and are always embedded in larger assemblies. An example for a large protein family containing parallel β-barrels is the triosephosphate isomerase (TIM) family (Figure 18.37). The β-barrel in their center is entirely hydrophobic on its outside. The hydrophobic surface is completely shielded from solvent by the amphipathic helices surrounding the β-barrel.

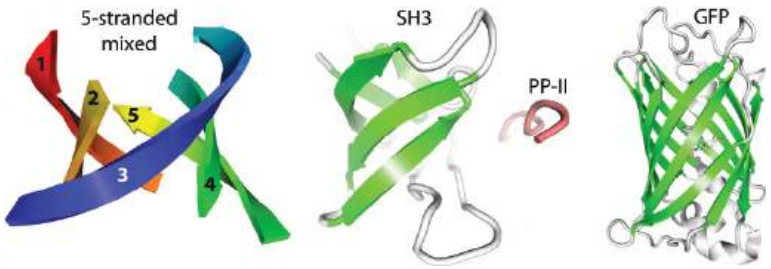


FIGURE 18.32: β-barrels in proteins. The smallest β-barrels are 5-stranded with four strands being antiparallel and the first and last strands (1 and 5) oriented parallel to each other. SH3 domains are examples of such small barrels that bind to Pro-rich sequences in the poly-Pro-II conformation (Figure 18.26). The green fluorescent protein (GFP) is an antiparallel 13-mer barrel housing a fluorophore made by post-translational modification of amino acids.

18.3.3 Reverse turns

Reverse turns change the direction of a polypeptide by up to 180° over as little as three or four residues. Also known as *tight turns*, β-*turns*, or simply *turns*, reverse turns are the most common type of non-repetitive structure in proteins. Turns can make up anywhere between 24% and 45%

of the entire protein structure. They are responsible for the globular forms that proteins adopt, and frequently the sites where nascent polypeptide chains emerging from the ribosome begin folding into their three-dimensional structures. Most turns have predominantly hydrophilic side chains and are located at the surface of proteins (Table 18.5).

Most turn geometries allow formation of a hydrogen bond between the first residue *i* and the *i*+3 residue. Reverse turns connecting adjacent antiparallel β-strands often form a hydrogen bond that is at the same time part of the β-sheet. However, a survey of turns in protein structures found that only about half of all turns are hydrogen-bonded at all, indicating that the presence of a hydrogen bond in turns may be a consequence of turn formation, not a prerequisite. If the hydrogen bond is formed, it often is non-linear, and does not contribute much to the stability of the turn. In agreement with this observation and from the study of turns in model peptides, most turns are not stable on their own but gain stability within the framework of the whole protein.

Turns are classified into at least ten types according to distinct structural features, including torsion angles, interactions within the turn, and steric requirements. The vast majority of turns in protein structures belongs to the six types I, I', II, II', VIa, and VIb (Figure 18.33), and we will focus on those in more detail. Other classes of turns (termed type III, IV, VII, and γ) are slight structural variations of these themes that are much less frequent and therefore not further discussed here.

Type I and type II turns fold the peptide chain around an almost square corner. Both turns bend the polypeptide chain in a way that in principle allows continuation of an antiparallel β-sheet. Type I and type II turns are related to one another by a 180° flip of the central peptide unit around the main-chain axis. Turns of types I' and II' have backbone conformations that are mirror images of types I and II: the ϕ/ψ angles of the central two amino acids are inverse to those of type I and II (Table 18.8), changing the handedness of the turns as well. Type I' and II' turns are well-suited to connect an antiparallel β-sheet because their handedness aligns better with the right-handed twist of the β-strands in the sheet (Figure 18.31) than types I and II.

Turns of type VI have a *cis*-Pro at position *i*+2. There are two distinct conformations, type VIa and VIb. The Pro in type VIa has near α-helix backbone conformation while the Pro in type VIb has β-conformation. The curved "plane" made by the *cis*-Pro peptide group and the Pro side chain has a more "concave" orientation relative to the overall curve of the turn in type VIa but is more "convex" in type VIb (Figure 18.33). Whereas type VIa is typically hydrogen-bonded, type VIb is usually not. Pro is also prominent in other turns and fits well at position *i*+1 of types I and II turns, and at position *i*+2 of type II'.

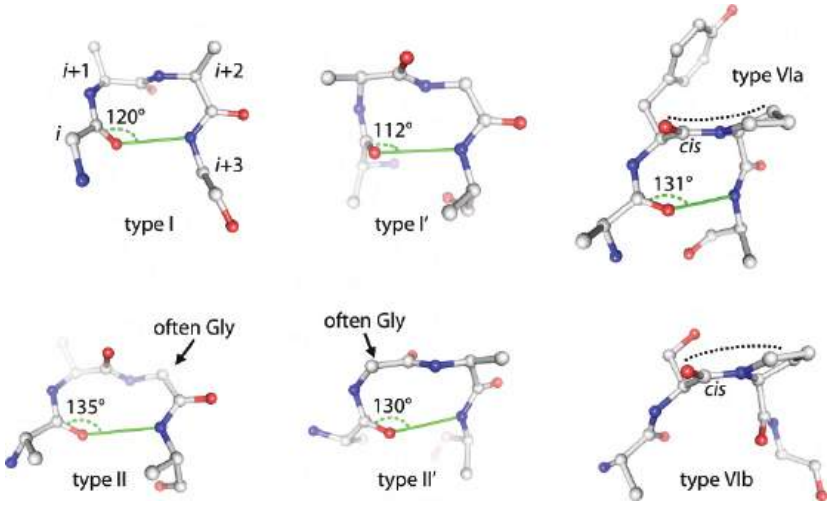


FIGURE 18.33: Frequent turns in protein structures. Hydrogen bonds are shown as green lines between residues *i* and *i*+3. The angle between the hydrogen bond donor (carbonyl) and the acceptor (NH) is also given. The hydrogen bond angles deviate significantly from the optimal geometry of 180°. In fact, hydrogen bonding is not a prerequisite for the existence of a turn. Type VI turns have a *cis*-Pro peptide bond. The subtypes VIa and VIb differ mainly by the main-chain conformation on both sides of the Pro. Type VIa is usually hydrogen-bonded, type VIb is not.

In contrast to α-helices and β-strands, reverse turns have different ϕ/ψ combinations for each residue. Most of these ϕ/ψ angles fall into the energetic minima of the Ramachandran plot for regular secondary structures (Figure 18.21). The central two residues, *i*+1 and *i*+2, can have torsion angles outside these regions (Figure 18.34). In fact, the main-chain torsion angles for these residues are diagnostic for the type of turn (Table 18.8).

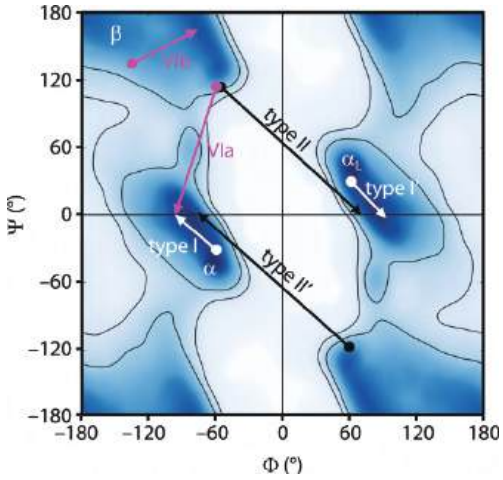


FIGURE 18.34: ϕ/ψ angle combinations in reverse turns. The Ramachandran plot for Gly shows more and shallower energy minima than those for the other canonical amino acids (Figure 18.21). The arrows start with a circle at the ϕ/ψ angles for residue *i*+1 and end at those for residue *i*+2. The Pro-containing type VIa and VIb turns are also shown (magenta arrow heads). Figure created with the programs Phenix and Molprobity.

TABLE 18.8:
ϕ/ψ combinations for the central two amino acid residues in turns.

turn type	ϕ(*i*+1)	ψ(*i*+1)	ϕ(*i*+2)	ψ(*i*+2)
I	−60	−30	−90	0
I'	60	30	90	0
II	−60	120	80	0
II'	60	−120	−80	0
VIa	−60	120	−90	0
VIb	−135	135	−75	160

In the description of the different types of turns, we have already seen that the tight reversal of the backbone in a turn poses steric constraints on the side chains. It is therefore not surprising that not all amino acid residues are equally well suited for a tight turn. We have encountered Gly and Pro as residues that nicely fit into the tight geometry of turns, and indeed Gly and Pro are the most frequent amino acids observed in turns. Following them in frequency are the small and hydrophilic residues Asn, Asp, and Ser. The high preference of Gly in type I', II, and II' is due to the rather strict dihedral angle restraints for the central positions (*i*+1 and *i*+2), which only Gly can satisfy. For instance, in the type II turn, the carbonyl oxygen of the central peptide group between residues *i*+1 and *i*+2 would come very close to a non-hydrogen side chain in position *i*+2, which is avoided when having Gly at this position. Indeed, 61% of type II turns have Gly at this position. A similar clash would exist between the NH-group of the central peptide group and a larger side chain in position *i*+1 in a type II' turn, which is why Gly is the preferred residue here. The side-chain ring of

Pro locks the ϕ angle at −60°, so its energy minima roughly correspond to those of the α-helix and the β-strand (Figure 18.21). In type VIa turns, the Pro has α-conformation while in type VIb, it has β-conformation. In summary, turns are sites of chain reversals in the folded protein, which due to their tightness and surface location require certain small and often hydrophilic side chains that can adopt special torsion angle combinations.

The positions of the secondary structure elements discussed in Sections 18.3.1–18.3.3 can be predicted from sequence information (Box 18.2). Such an assignment is useful for proteins of unknown structures, particularly in cases where protein folding algorithms (Section 20.3) struggle, such as flexible peptides, small proteins with post-translational modifications, or membrane proteins.

BOX 18.2: SECONDARY STRUCTURE PREDICTION.

We saw before that different amino acids have different propensities for α-helices, β-sheets, and turns (Table 18.5). This statistical information from proteins of known structure can be used to predict secondary structure elements purely from sequence analysis. The first algorithms used single sequences, but better results with > 70% accuracy were later obtained by using multiple sequence alignments (MSAs). MSAs outline conserved sequence regions in whole protein families, which are expected to have conserved secondary and tertiary structure as well. One of the first algorithms was the *Chou-Fasman method*, which assigns a probability for each amino acid to occur in helices, sheets, or turns. In a sliding window analysis, a window of 5–10 amino acids is shifted along the sequence, and the propensities for the secondary structures are tallied. Consistently high values for a particular secondary structure element indicate that the sequence may form this secondary structure. However, a single prediction is still not very accurate, and it is best to combine the results of several different algorithms such as JPRED or Pep2D. Different methods tend to agree on the position of helices, loops, and strands, but may predict their start and end differently.

Secondary structure prediction works much better for α-helices than for β-strands. The reason is that α-helices are characterized by local interactions between a continuous stretch of residues in a sequence, whereas the interaction partners for β-strands are far away in the protein sequence. Many α-helices on the surface of proteins are amphipathic, with one side of the helix contributing to the hydrophobic core of the protein and the other being solvent-exposed. Such helices can be recognized by two shifted *i*,*i*+4 patterns of hydrophobic and hydrophilic residues. Transmembrane helices may also be predicted: a membrane is ca. 2 nm thick, and the α-helix has a 540 pm pitch, so it takes 4–5 turns or 16–20 residues with hydrophobic side chains for the shortest α-helix to traverse the membrane. Figure 18.35 shows the secondary structure prediction of bacteriorhodopsin together with the hydropathicity data from Figure 18.17. Combination of these complementary methods can increase the accuracy of the prediction of transmembrane regions compared to either method alone.

Information on possible secondary structure elements can sometimes be obtained by direct inspection of the sequence. For instance, while Pro and Gly are both disfavored in shorter strands and helices, we have seen in the previous Section (18.3.3) that Pro-Gly or Gly-Pro sequences can be diagnostic for a turn.

(*Continued*)

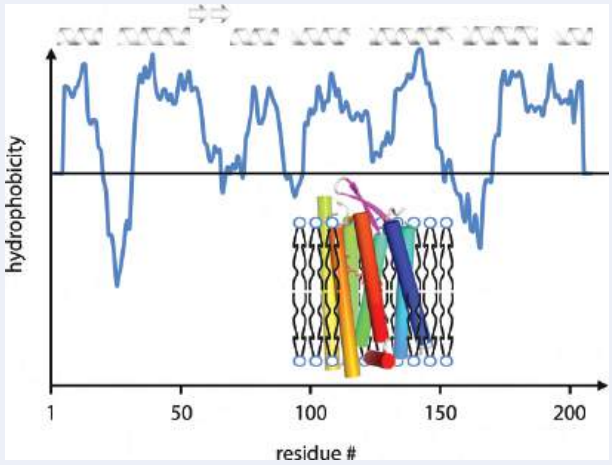


FIGURE 18.35: Secondary structure prediction. The structure of *Haloarcula marismortui* bacteriorhodopsin (PDB-ID 4pxk; same sequence) is shown as an inset with a sketched membrane. The helix segments present in the structure are indicated above the hydrophobicity data from Figure 18.17. All secondary structure elements were correctly predicted by JPRED (top), including the two short β-strands (magenta in the structure). The predicted helices align roughly, but not perfectly, with the hydropathicity data.

With this, we conclude our excursion into secondary structure prediction, and turn from secondary structure elements to the next level of protein structure formation: tertiary structure.

18.3.4 Protein domains and tertiary structure

Tertiary structure formation follows the same principles that govern secondary structures, maximization of attractive interactions, and minimization of steric clashes and hydrophobic surface exposed to water. The tertiary structure is the three-dimensional conformation of a single polypeptide chain. Tertiary structure formation begins with the connection of several secondary structure elements into super-secondary structures. Super-secondary structures become further stabilized when assembling into larger protein *domains* (Figure 18.36). Domains are the principal building blocks of proteins. They are compact, globular entities stabilized by tertiary interactions between residues far apart in sequence. The strict definition of a protein domain is an *autonomous folding unit*, *i.e.* a sequence of amino acids with the intrinsic ability to fold into a stable, compact structure. Domains are usually associated with a specific function. In biology, a more lenient definition of a domain is any more or less globular part of a protein that has a defined function, for instance a catalytic activity or a binding property.

Examples of small domains that are autonomous folding units are WW, SH2, PH (Figure 18.36), and SH3 domains (Figure 18.32). Domains may become very large, especially when they assemble from repeating super-secondary structures. *Repeat proteins* contain multiple copies of the same small super-secondary structure, *e.g.* a helix-loop-helix motif or a strand-loop-helix motif, and they display a broad spectrum of shapes, ranging from planar, crescent-shaped, or circular forms to helical solenoids (Figure 18.36). Many proteins consist of multiple domains that can further interact with each other through additional tertiary interactions, which in this case are termed inter-domain interactions.

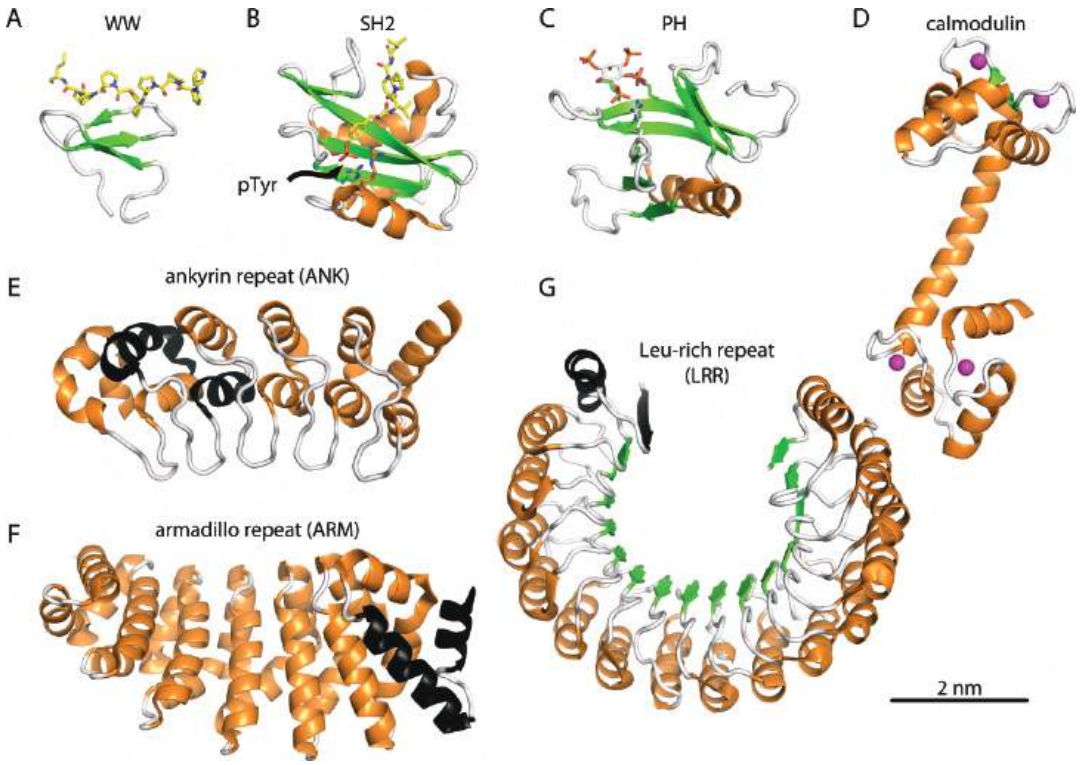


FIGURE 18.36: Super-secondary structures in small protein domains and repeat proteins. A: Similar to the SH3 domain (Figure 18.32), the WW domain recognizes Pro-rich peptides, mediating protein-protein interactions. It is a tiny protein domain of only 40 residues with a three-stranded β-sheet that contains two signature Trp residues (PDB-ID 1upr). Some WW domains bind peptides containing phosphoserine. B: SH2 domains are phosphotyrosine (pTyr) binding domains. Phosphorylation of a substrate protein by a tyrosine kinase triggers SH2 domain binding, often in the context of cellular signaling (PDB-ID 1h9o). C: PH domains are binding modules for phosphoinositide lipids (Table 18.10). Proteins with PH domains often function as adapters that bind and tether other proteins to membranes. The example shown (PDB-ID 1upr) is in complex with inositol 1,3,4,5-tetrakis phosphate, the head group of an important class of phosphoinositide lipids. D: The Ca^{2+}-binding protein calmodulin has several EF-hand motifs. The EF-hand is the most common Ca^{2+}-binding motif and consists of two short α-helices connected by a loop that binds the ion (PDB-ID 1exr). E-G: The repeating super-secondary structures in three repeat proteins are marked in black. E: The Ankyrin repeat (ca. 33 residues) has two α-helices separated by a long loop (PDB-ID 2rfm). F: The slightly longer (ca. 40 residues) Armadillo repeat consist of two α-helices separated by a turn (PDB-ID 4plr). G: Conserved Leu residues between an α-helix and a β-strand coined the name for the Leu-rich repeat (PDB-ID 1z7x).

Based on their secondary structure content, protein domains can be classified into all-α, α/β, and all-β (Figure 18.36). All-α domains only contain helices connected by loops. α/β-domains contain a mixture of β-barrels and/or twisted β-sheets along with α-helices that pack against the β-structures. All-β domains exclusively contain β-barrels and β-*sandwiches*, all of which are usually formed by antiparallel β-sheets. The β-sandwich is a variant of the β-barrel where two β-sheets lie on top of each other in an almost perpendicular orientation so the outer strands do not interact as intimately as they do in barrels (Figure 18.37). *Ig-* or *immunoglobulin* domains (Figure 18.37) are examples of very stable β-sandwiches, made from two Greek key motifs that pack against each other and that are covalently linked by a disulfide bond (Section 18.1.4.5).

The topology of secondary structure elements in a domain determines the *fold* of the domain. The number of different folds that proteins can adopt is finite. There are about 1300 fundamentally different structures a polypeptide chain can adopt, and these folds are used to classify any protein structure (Box 18.3). About ten different folds appear very successful during evolution and occur in 30% of all proteins. Some of these *superfolds*, such as the *Rossmann fold* and the *globin fold*, are tied to specific biochemical functions, while other superfolds, such as the Ig-fold, TIM-barrels, and β-propellers, are quite variable in function.

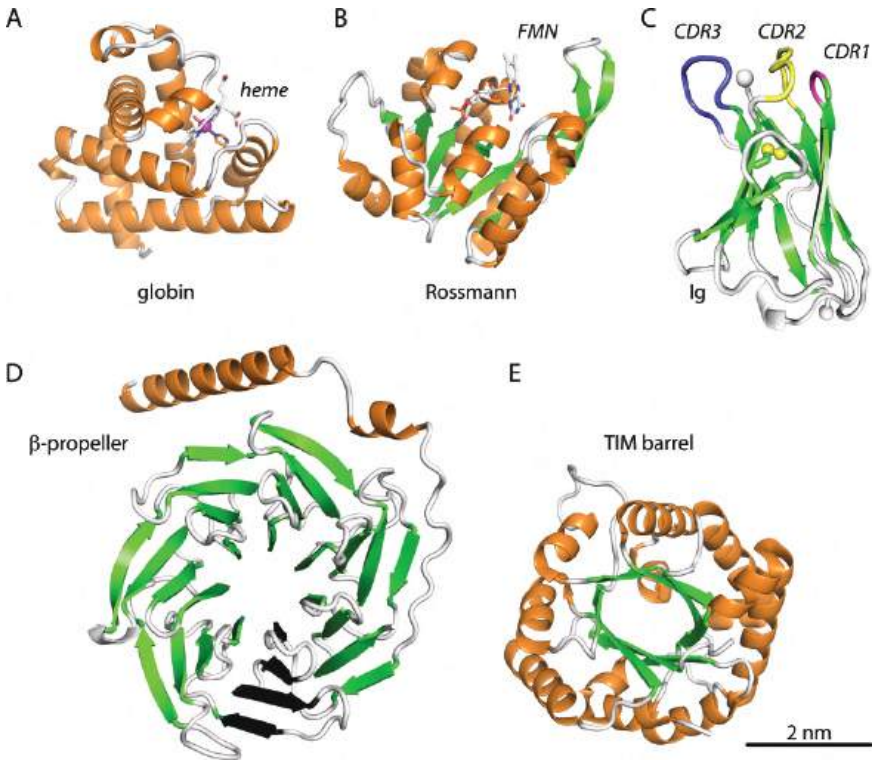


FIGURE 18.37: Recurring protein folds. A: The globin fold is made entirely of α-helices that form a hydrophobic pocket for a heme group (PDB-ID 1a6k). B: The Rossmann fold is a mononucleotide-binding module (PDB-ID 1g5q). C: The immunoglobulin (Ig) domain (taken from PDB-ID 1igt) is a small and very stable β-sandwich. D: β-propellers are circular assemblies of small antiparallel β-sheets. The example shown here is the β-subunit of a heterotrimeric G-protein, which serves as a binding platform for the α- and γ-subunits (PDB-ID 1got). E: In TIM-barrel proteins, the central parallel β-barrel is stabilized by the crossover α-helices on the outside (PDB-ID 1tim).

The Rossmann fold (Figure 18.37) is a mononucleotide-binding motif that is usually embedded in a larger protein. Its structure is a sequence of βαβ elements, which generates a central parallel β-sheet with amphipathic α-helices as crossovers. Superfolds have a modular character: enzymes with two Rossmann folds bind dinucleotides, such as the cofactors FAD and NAD^+ (Figure 21.13). The globin fold (Figure 18.37) encompasses eight α-helices that form a pocket for heme to bind oxygen in the transport and storage proteins myoglobin and hemoglobin, and in cytochromes, which are important de-toxifying enzymes in the liver. The Ig-fold, a disulfide-bonded β-sandwich (Section 18.1.4.5), is a ubiquitous structural unit in cell surface receptors and also present in all antibodies, including their antigen-recognition site. The β–propeller is a superfold that has a circular arrangement of antiparallel β-sheets connected by loops or other secondary structures. β–propellers can vary widely in form and function. The number of blades in the propeller is between 4 and 9, although 5- and 6-bladed propellers are most common. β–propellers are present in enzymes, structural proteins, and proteins that bind to other proteins. A prominent example of a β-barrel is the triosephosphate isomerase (TIM) barrel, an 8-stranded parallel $(\beta\alpha)_8$ fold that is found in about 10% of all enzymes, including such diverse functionalities as hydrolases, lyases, isomerases, and oxidoreductases.

We now turn from single-domain to multi-domain proteins. Larger (> 50 kDa) proteins tend to contain several domains, organized like pearls on a string with linkers connecting them. Depending on the length and flexibility of the linkers, the domains may or may not interact with each other. If the linkers are rather short, the domains may functionally cross-talk. If the linkers are long and/or flexible, the individual domains connected by the linkers can adopt very different juxtapositions relative to one another. Very flexible linkers often contain sequences rich in Gly, Pro, Ser,

or Gln. Flexibly linked domains may not necessarily functionally interact, but the domain movements enable large conformational changes of the protein, which is important for function. Titin is an extreme example of a multi-domain protein where the domains apparently do not functionally cross-talk. This vertebrate muscle-organizing protein has the largest single polypeptide chain in humans with 27000–38000 residues depending on the titin isoform. Among others, titin harbors a kinase domain, 152 immunoglobulin domains, and 132 fibronectin-like domains (β-sandwiches similar to immunoglobulin domains). While it is often observed that individual domains can be isolated from multi-domain proteins and are stable in solution, for many other proteins the linker is important for domain stabilization and is in fact part of the domain.

BOX 18.3: DOMAIN ALIGNMENT AND FOLD RECOGNITION.

One of the first questions when a new protein structure has been determined by any structural biology method is whether it has known structural homologs or whether a novel topology (fold) has been discovered. For this purpose, the coordinates are often submitted to the Distance mAtrix aLIgnment (DALI) server. The DALI method is based on the fact that atoms in similar structures have similar distances. Although the sequences may be entirely unrelated, this distance similarity criterion surprisingly often identifies a common fold. Indeed, the number of novel folds added to the protein databank (PDB) in recent years is very small. The Class, Architecture, Topology/Homologous superfamily (CATH) database classifies protein structures according to secondary structure content, structural similarity in the absence of sequence similarity, fold, and evolutionary relationship. According to the CATH 2024 release, there are only 1739 different folds. A similar database that uses purely structural considerations is the Structural Classification Of Proteins (SCOP2) database. In June 2022 this database defined 1562 folds, with only 26 new folds added between 2017 and 2021. There is a fair chance that any new protein sequence is already represented by a known structure. Before the advent of accurate *ab initio* protein folding algorithms, techniques known as *fold recognition* and *threading* were used to find novel folds (Section 20.3).

18.3.5 Quaternary structure

When several polypeptide chains assemble into higher-order oligomers they form a *quaternary structure*. The enthalpic term for the formation of quaternary structures is driven by a favorable enthalpic term, caused by the attractive interactions formed at the protein-protein interface. The entropic term contains both favorable and unfavorable contributions. The association of individual subunits into the oligomer as well as restrictions in conformational freedom of each polypeptide chain lead to a decrease in entropy. At the same time, water is released from the protein-protein interface, leading to an increase in entropy. Oligomer formation is therefore a good example of *enthalpy-entropy compensation* (Section 1.6.1). The relative contributions of enthalpy and entropy determine the position of the monomer-oligomer equilibrium. We will see more examples of enthalpy-entropy compensation throughout this Section.

Oligomerization can be reversible or irreversible. Irreversible oligomers may be stabilized by covalent crosslinks such as disulfide bonds. If the assembly is reversible, the oligomer is normally in dynamic equilibrium with its constituent monomers.

Oligomerization can take place between several copies of the same protein (homo-oligomerization) or between one or more copies of different proteins (hetero-oligomerization). A survey of crystal structures of oligomeric proteins (Table 18.9) shows that homo-oligomers are more frequent than hetero-oligomers. In addition, small stoichiometries are preferred over large stoichiometries, and even-numbered components are preferred over odd-numbered components.

TABLE 18.9:
Stoichiometries of the most frequent oligomers.

homo α_n	2 (32)	3 (3.8)	4 (8.2)	5 (0.40)	6 (2.6)	7 (0.12)
homo α_n (ctd.)	8 (1.0)	10 (0.34)	12 (0.56)	24 (0.17)	48 (0.13)	60 (0.16)
hetero $\alpha_n\beta_m$	1,1 (4.6)	2,1 (0.48)	2,2 (1.7)	3,3 (0.29)	4,4 (0.22)	4,2 (0.16)

Data are from the protein databank (PDB) as of January 2016 (35460 crystal structures with < 90% sequence identity). Newer releases of the PDB unfortunately do not provide this kind of analysis any more, but there is no reason to suspect a large change in the oligomer frequencies. Only oligomers > 0.1% frequency are listed. Numbers in parentheses are percentages with the values for *n* and *m* (in hetero-oligomers) in front. For comparison, there are 46% monomers. Heterotrimers of the form αβγ constitute 0.89%, and αβγδ heterotetramers contribute 0.28% to all structures.

There is a weak correlation between the molecular mass of a protein and its tendency to oligomerize: monomeric proteins generally have rather small molecular masses of around 30 kDa, whereas larger proteins > 50 kDa molecular mass tend to form oligomers. Assembly into a quaternary structure is a general strategy in biology to add protein functionality, increase protein stability, and enable regulation of binding and function through cooperative effects (Sections 2.4, 3.4.3, 13.5). Oligomers frequently undergo cycles of large conformational changes that may not be accessible to the smaller monomers.

18.3.5.1 Homo-oligomers

Homo-oligomers are assembled from several copies of the same monomer. They often display internal symmetry (Figure 18.38) and incorporate two-fold, three-fold, etc. rotation axes that relate one monomer to the other(s). Such a symmetric arrangement maximizes the number of interactions between the monomers and is therefore energetically favored compared to the monomers.

The simplest and most frequent oligomer is the homodimer, found in 32% of all reported crystal structures. Among them is glutathione S-transferase (GST), which is often used as an N-terminal fusion (tag) to purify a protein of interest by affinity chromatography on glutathione sepharose. An example of a trimer is the 2-keto-3-deoxy-6-phosphogluconate aldolase, which forms a planar triangle. While there is only one way to arrange monomers into symmetric dimers and trimers, four monomers can be arranged symmetrically in different ways. A tetramer can either be planar, like the Borna disease virus nucleoprotein, or adopt the form of a tetrahedron, with the monomers sitting at the four corners, such as lactate dehydrogenase (Figure 18.38). Hexamers can be planar six-membered rings or dimers of trimers or trimers of dimers. Octamers can be planar rings, dimers of tetramers, or tetramers of dimers, and so on for higher-order oligomers. For uneven numbers of monomers, the most prominent arrangement is a planar ring or a helix (a ring with a pitch). GroEL, for example, is a chaperone composed of two regular seven-membered rings stacked on top of each other that enclose a cavity in which proteins can fold (Figure 18.38 and Figure 18.39). When tubulin monomers assemble into filaments, they form an "infinite" helix with 13 monomers per turn. Oligomers composed of many monomers can also from hollow shapes that function as containers. An example is apoferritin (Figure 18.38), which forms a hollow sphere made from 24 identical subunits to store iron in. The concept of storage by formation of hollow spheres is also used by many viruses for their genomic material (see Section 18.3.5.2).

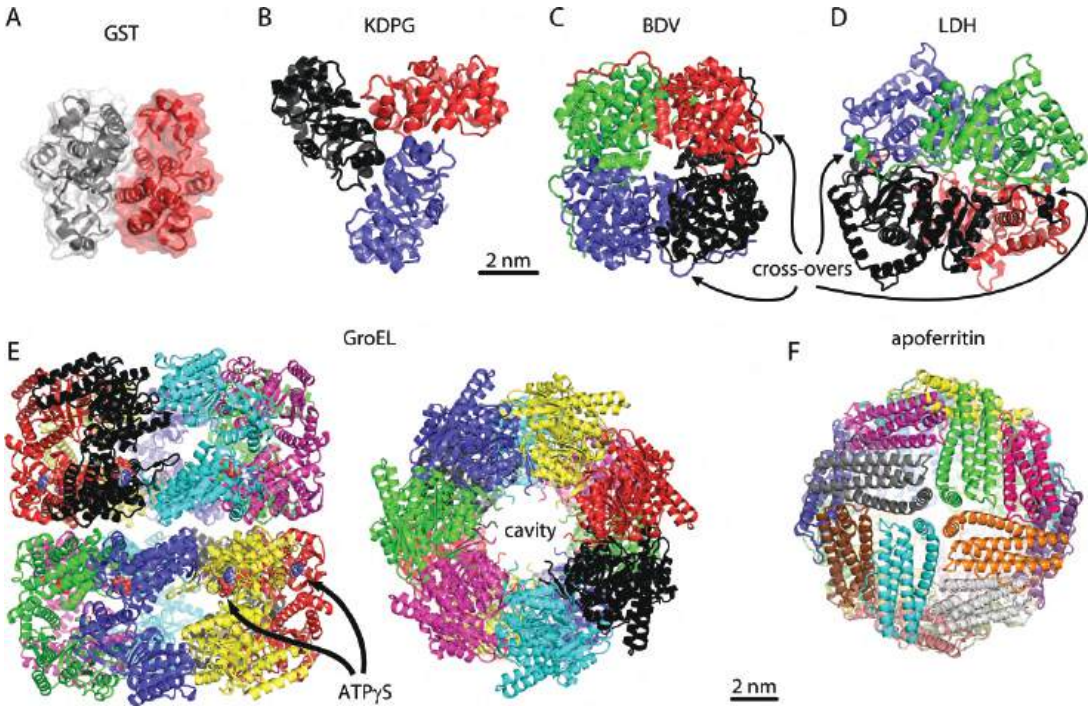


FIGURE 18.38: Quaternary structure of selected homo-oligomers. For all examples but the side-view in (E) the rotation axis points toward the reader. A: *Schistosoma mansoni* glutathione S-transferase (GST; PDB-ID 1u31) is a symmetric dimer. The subunits are colored black and red with transparent surfaces to highlight the interface. B: *E. coli* 2-keto-3-deoxy-6-phosphogluconate aldolase (KDPG; PDB-ID 1fq0) is a trimer. C: The Borna disease virus nucleoprotein (BDV; PDB-ID 1n93) is a planar tetramer. D: By contrast, the human muscle L-lactate dehydrogenase (LDH; PDB-ID 4ojn) forms a tetramer in the shape of a tetrahedron. There are large crossovers from one monomer to the other, which helps to stabilize the oligomer. E: The *E. coli* chaperone GroEL is a large 14-mer composed of two seven-membered rings stacked on top of each other. This structure has the non-hydrolyzable ATP analog ATPγS bound (PDB-ID 1sx3). Two orientations rotated 90° from each other are shown. F: Apoferritin is a hollow sphere made of 24 α-helical subunits (PDB-ID 2w0o).

18.3.5.2 Hetero-oligomers

Hetero-oligomers are formed by association of different subunits. The formation of hetero-oligomers follows the same principles relating to the structure and stability as we discussed for homo-oligomers. Hetero-oligomers can be symmetric, *pseudo-symmetric*, or asymmetric. Pseudo-symmetric shapes are possible if the subunits are of similar structure. Hemoglobin is an example of a pseudo-symmetric $\alpha_2\beta_2$ heterotetramer, formed by α- and β-subunits with similar, but not identical structures (Figure 18.39). The $\alpha_6\beta_6$ heterohexamer of aspartate transcarbamoylase is composed of two α_3-trimers and three β_2-dimers. Pseudo-symmetry can repeat many times to form the hollow spheres of capsids that store viral genomes. In foot-and-mouth disease virus (Figure 18.39), three similar subunits form a trimer, which further assembles into the capsid. More examples of hetero-oligomers with at least some pseudo-symmetry include lactose synthase (αβ), tryptophan synthase ($\alpha_2\beta_2$), cholera toxin ($\alpha\beta_5$), ribulose bisphosphate carboxylase ($\alpha_8\beta_8$), troponin (αβγ), phosphorylase kinase ($\alpha_4\beta_4\gamma_5\delta_4$), RNA polymerase ($\alpha_2\beta\beta'$), F_1 ATPase ($\alpha_3\beta_3\gamma\delta\varepsilon$), and antibodies.

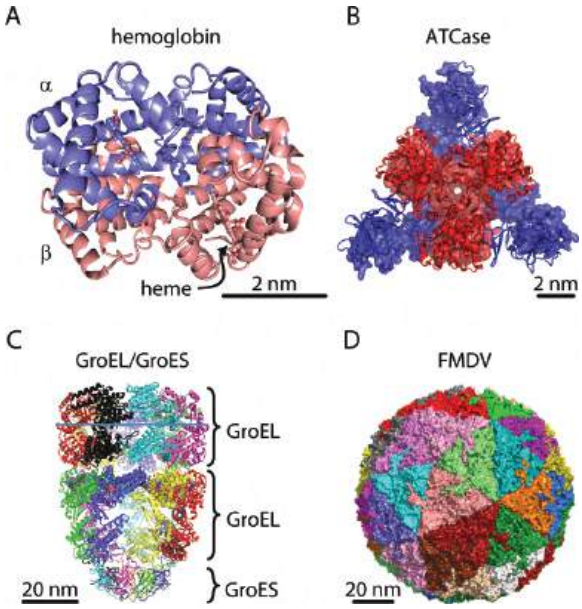


FIGURE 18.39: Quaternary structure of selected hetero-oligomers. A: Fetal deoxy-hemoglobin is an $\alpha_2\gamma_2$ heterotetramer (blue and pink; PDB-ID 2hhb), B: Aspartate transcarbamoylase (ATCase; PDB-ID 1q95) is an $\alpha_6\beta_6$ heterohexamer. The catalytic α_3 trimers are colored red and the regulatory β_2 dimers are shown in blue. C: The small GroES homoheptamer attaches to the GroEL ring and induces large conformational changes in GroEL. D: Foot-and-mouth disease virus (FMDV; PDB-ID 1bbt) is a hollow icosahedral sphere that contains the single-stranded RNA genome of the virus. The icosahedron is assembled from three similar subunits (VP1, VP2, and VP3) that form trimers (colored), which further assemble into the viral capsid.

Hetero-oligomers composed of different monomers lack internal symmetry. Many complexes between RNA and proteins belong to this class, for instance ribosomes and spliceosomes (Figure 19.40), which contain dozens of different and structurally unrelated protein subunits that form intricate networks of non-covalent bonds across molecular interfaces. In Section 18.3.6, we will discuss the nature of protein-protein interactions that form the interfaces of macromolecular complexes. Protein-nucleic acid interactions will be treated in the context of nucleic acids, in Section 19.5.4.1.

18.3.6 Protein-protein interfaces

Protein-protein interfaces occur in domain-domain interactions, in binary interactions between proteins, and in interactions between individual subunits of oligomers. In all cases, assembly is driven by the maximization of non-covalent interactions, and minimization of exposed hydrophobic surfaces. The main characteristics of the interacting molecular surfaces are shape complementarity, biophysical complementarity, and in some cases a few embedded water molecules (Figure 18.40).

High shape complementarity maximizes the number of contacts in a complex by avoiding cavities within the interface. Cavities in the interior of proteins and protein complexes are usually destabilizing, we will revisit this in the context of protein folding in Section 18.5. Biophysical complementarity in addition refers to the spatial arrangement of interacting groups such that a maximum number of attractive interactions can be formed across the interface while the number of unsatisfied hydrogen-bonding valences is minimized. Due to their distinct directionality, hydrogen bonds are important for the specificity of the interaction. Upon complex formation, most water molecules and ions that cover the surface of the isolated binding partners in solution are released from the surfaces that form the interaction interface, leading to a favorable entropy change. High-affinity complexes usually have no or just a few water molecules embedded in their interface. If water molecules are locked in the interface, they tend to form a maximum number of hydrogen bonds. Polar or charged groups from the binding partner must form hydrogen bonds or

salt bridges across the interface to avoid energetically unfavorable burial of polar atoms in the non-polar environment of the complex.

18.3.6.1 Surface complementarity and buried surface area

High-affinity macromolecular complexes usually have an extensive and hydrophobic interface due to the hydrophobic effect that disfavors solvent-exposure of these surfaces. Moreover, surfaces composing the interface of high-affinity complexes have high shape complementarity. The better the shapes of the interacting surfaces fit together, the more attractive interactions may be formed. Shape complementarity of a complex can be calculated and expressed as a number between zero and unity, termed the shape complementarity coefficient S_c. A value of zero means no fit at all and a value of unity signifies perfect complementary (Figure 18.40). Typical values for protein-protein complexes range from 0.5–0.9. In the absence of affinity information, S_c calculated from crystal structures of several related protein-protein complexes may help in judging their relative affinity in solution. When analyzing the macromolecular interactions seen in crystals (crystal contacts; Section 24.2), S_c may also help in distinguishing a biologically relevant interaction surface from an irrelevant one.

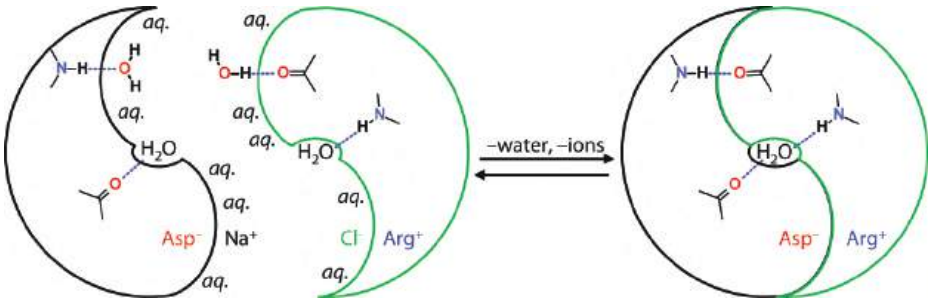


FIGURE 18.40: Biophysical and shape complementarity of macromolecular interfaces. Before subunit association, charges are screened by counter ions (Na^+ and Cl^-) and the surfaces are covered with water molecules, symbolized by "*aq.*". Upon complex formation, most solvent molecules are displaced from the interface. High-affinity complexes bury large interfaces and have high shape complementarity. Occasionally, some water molecules are trapped in the interface and become an integral part of the complex by forming a large number of hydrogen bonds.

On average, proteins bury 10–20% of their surface in a complex interface. In addition to the S_c, the *buried surface area* (BSA) is a rough indicator for protein complex stability. Generally, the larger the BSA the stronger the interaction. If the structure of the complex is known, the BSA can be calculated from the difference in *accessible surface areas* (ASA) of the isolated binding partners and the complex. The ASAs are determined by rolling a sphere with the radius of a water molecule (170 pm) over the surface, and tallying the contact area. The BSA is then the sum of the ASAs of the constituent proteins minus the ASA of the complex. Common sizes of interaction surfaces are in the range of 10–55 nm^2. However, there is considerable variety and much smaller surface areas have been observed for stable complexes.

A statistical analysis of surfaces in protein-protein complexes has shown that complex interaction surfaces of less than 20 nm^2 are usually one continuous patch on the surface of each binding partner. In contrast, larger interfaces tend to be divided into several contacting patches with at least one dominating patch of larger size. The largest patch, or primary interface, is similar in size and biophysical properties to that present in the single-patch interfaces of smaller complexes. Surrounding the primary interface are smaller patches, or secondary interfaces. The contributions of the different residues to complex stability in a patch vary. There are interaction *hot spots* at the center of an interface patch, where usually a set of buried atoms is responsible for the majority of the interactions. This hot spot is surrounded by atoms at the perimeter of the patch that contribute less to the interaction and that may even be solvent-accessible.

18.3.6.2 Energetics of interactions at protein-protein interfaces

Protein-protein interactions involve a large number of individual molecular interactions. Each individual interaction in a protein-protein interface contributes only little to complex stability, but the small values can add up to considerable stabilization energies. For example, while the insulin dimer has a K_d value of only 10^{-5} M, the strong trypsin/trypsin-inhibitor interaction has a K_d value of 10^{-13} M, corresponding to interaction energies between –28 and –74 kJ mol^{-1}, respectively, at 25°C. About ⅔ of the interaction surfaces is non-polar, and hydrophobic interactions have the largest contribution to complex stability. Occasionally, buried salt bridges can also strongly contribute to binding. Buried salt bridges are stronger than solvent-accessible salt bridges because of the lower dielectric constant in the interface compared to bulk solvent (Section 17.2.2.1). Proteins from thermophilic organisms often contain more salt bridges across the subunit interface than their mesophilic counterparts as a general strategy to stabilize oligomers at the elevated temperatures of their surroundings. Buried hydrogen bonds are also stronger than surface-exposed hydrogen bonds. On complex formation, hydrogen bonds with water in the monomeric subunits are replaced by hydrogen bonds between the subunits in the complex. Due to the lower relative dielectric constant in interfaces compared to bulk solvent ($\varepsilon_r \approx 80$), the hydrogen bonds formed at the interface are somewhat more stable than those broken at the surface. Thus, they contribute to some extent to the overall oligomer stability, but contribute less to complex stability than buried ion pairs.

The energetic contribution of individual side chains to complex stability can be determined by measuring the effect of mutagenesis at the subunit interface on complex stability. Similarly, the stabilities of natural mutations can be compared. Removal of a single interaction from a protein-protein interface can either destabilize or stabilize an oligomer. For example, a natural mutation in the β-chain of hemoglobin, Tyr35Phe, removes a hydrogen bond to Asp126 in the α-chain. People carrying this hemoglobin variant "Philly" suffer from mild anemia because loss of the hydrogen bond leaves a destabilized $\alpha_2\beta_2$ tetramer, leading to its dissociation into monomers.

The attractive interactions at the protein-protein interface manifest as a favorable enthalpic contribution to complex stability. Again, the favorable enthalpic term is counteracted by unfavorable entropic contributions on complex formation, caused by the confinement of the interaction partners in a single particle, conformational restrictions at the interface, and trapping of water molecules. The energetics of protein-protein complex formation is another example of enthalpy-entropy compensation.

We have seen that the water molecules bound to the binding partners prior to complex formation are an important factor for the contribution of polar interactions to complex stability. In the following, we will again see that water plays an even more important role for the interaction between non-polar groups.

18.3.6.3 Role of water – the hydrophobic effect

Water as the solvent will cover all exposed surfaces of a protein, with two important consequences. First, water will saturate all hydrogen bond donors and acceptors present on protein surfaces. Second, water will form ordered structures known as *clathrates* around hydrophobic surfaces such as the aliphatic or aromatic amino acid side chains of a protein, the bases of nucleic acids, or the hydrophobic groups of a ligand. While water forms four hydrogen bonds in ice and an average of 3.4 hydrogen bonds in solution, only about three hydrogen bonds can form on hydrophobic surfaces because part of the water molecule engages in van der Waals interactions with the hydrophobic groups (Figure 18.41). The hydrophobic surfaces also restrict the orientation and mobility of water, which reduces its entropy compared to bulk solvent. The resulting clathrate structures are the best compromise that water can achieve when bound to hydrophobic surfaces.

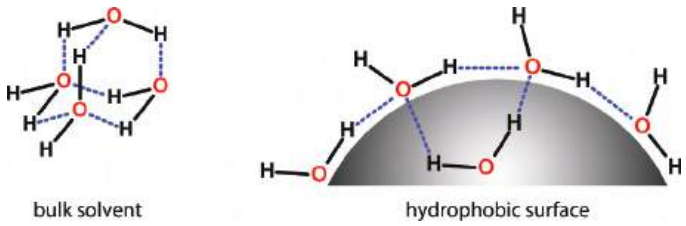


FIGURE 18.41: Bulk water and water at hydrophobic surfaces. The schematic drawings show possible hydrogen bonds as dashed blue lines. On hydrophobic surfaces fewer hydrogen bonds are possible between water molecules compared to bulk solvent.

Release of water from hydrophobic surfaces both increases the entropy of water and provides a favorable enthalpic contribution due to increased hydrogen bonding in bulk water. The positive entropy change contributed by a single water molecule can amount to 40–70 J mol^{-1} K^{-1}. This is a manifestation of the *hydrophobic effect*, which becomes important whenever hydrophobic groups are solvated by water.

The hydrophobic effect describes the tendency of molecules to exclude water from non-polar surfaces. The molecular reason that drives hydrophobic groups out of the aqueous phase is the increased interaction of water molecules in bulk solvent compared to when water is bound to hydrophobic surfaces. Bulk water is organized into networks of hydrogen bonds, which rapidly and constantly rearrange. When small, non-polar molecules such as noble gases, acetylene, or methyl sulfide are introduced into water, they disrupt some of these networks and dissolve all to about the same small extent by forming favorable van der Waals interactions with water molecules. The water molecules bound to the non-polar surface engage in fewer hydrogen bonds among each other, but these hydrogen bonds are slightly stronger compared to bulk solvent because solute-bound water molecules are less mobile. At low temperatures, the sum of van der Waals interactions and hydrogen bonds cause the overall ΔH to be slightly negative. However, at the same time several layers of water molecules become highly organized as cage-like clathrate structures around the non-polar surfaces. This ordering corresponds to a large and negative change in entropy ΔS, and a concomitant increase in heat capacity. Dissolving hydrophobic molecules in water is therefore entropically strongly disfavored, and the overall free energy change ΔG is positive: $-T \cdot \Delta S \gg \Delta H$, so therefore $\Delta G > 0$. At higher concentrations of the non-polar solute, its hydrophobic groups tend to interact with each other rather than with water. The switch in predominant interactions minimizes the decrease in entropy: the solute forms non-polar "islands" or droplets in a process known as *phase separation*. In biology, the exclusion of water from hydrophobic surfaces is the main driving force for protein folding, protein-protein and protein-ligand interactions, base stacking in nucleic acids, formation of membranes and vesicles by lipids, and the insertion of membrane proteins into membranes.

While polar interactions become weaker with increasing temperature, the temperature dependence of the hydrophobic effect is different. The hydrophobic effect first increases with temperature, which leads to a reduced solubility of hydrophobic solutes in water. With increasing temperatures, the water molecules fixed in the clathrate structures become increasingly mobile and interact less with each other. As a consequence, at high temperature both the entropic penalty for dissolving a non-polar molecule in water and the favorable change in enthalpy decrease. In other words, the entropic term becomes less unfavorable, and the enthalpic term becomes less favorable (enthalpy-entropy compensation). Because of the different temperature dependence of ΔH and ΔS (eq. 1.69, eq. 1.92), ΔG and thus the hydrophobic effect reach a maximum at the temperature where ΔS becomes zero (see Figure 1.26). At higher temperatures, ΔS becomes positive, but is now overcompensated by the large and positive ΔH, leading to a decrease of ΔG for transferring a non-polar molecule to water, and a decrease in the hydrophobic effect. This decrease of the hydrophobic effect occurs at very high, non-physiological temperatures (> 100°C).

The hydrophobic effect can be quantified by measuring the free energy of transfer of a hydrophobic molecule from a reference state (*e.g.* octanol, approximating the hydrophobic interior of a protein) into a hydrophilic environment (*e.g.* water). To a first approximation such transfer energies are additive and molecules can be conceptually broken down into individual contributions from their constituent parts (phenyl groups, amide groups, etc.). For instance, the transfer of methylene groups from the hydrophobic interior of a protein into bulk solvent has been estimated to 3.3 kJ mol^{-1} per methylene. Conversely, complete isolation of water from bulk solvent into a non-polar environment such as a hydrophobic cavity in a protein costs about 26 kJ mol^{-1} due to breaking of hydrogen bonds and the large positive entropy change.

18.3.7 Protein-ligand interactions

Ligand binding to proteins and DNA (Section 19.5.4.2) is governed by the same principles as protein-protein interactions, namely shape and biophysical complementarity. Shape complementarity maximizes the number of overall contacts and biophysical complementarity ensures productive polar interactions such as hydrogen bonds and ionic interactions. Although the extent of the interface is typically smaller in protein-ligand interactions than in protein-protein interactions, high-affinity ligands pack about as tightly to the protein as the packing of side chains in high-affinity protein-protein interfaces and the hydrophobic core of a protein, leaving little void space. Similar to protein-protein interactions, the main energetic driving force for protein-ligand interactions is the release of water from hydrophobic surfaces that become buried in the complex, another occurrence of the hydrophobic effect. Hydrogen bond donors and acceptors of ligands are paired with those of the protein, and the pronounced directionality of the hydrogen bonds formed imposes selectivity for the ligand. Hydrogen bonds between proteins and ligands can also be mediated by water molecules, in which case water becomes an integral part of the ligand binding site. If charges are present on the ligand, such as in nucleotides, they are usually oriented toward the solvent. Charges that are buried in the interface are neutralized by counter charges from metal ions or side chains, or by the partial charge of helix dipoles and peptide groups. In some cases, charges are efficiently solvated by a network of hydrogen bonds with oriented peptide dipoles that often connects to a nearby opposite charge. The charges are essentially distributed over a larger surface (dispersed) by the network. Another way to dissipate charges is to change the pK_A values of the ligand by the protein environment such that the ligand is bound in its neutral form.

All of these ligand-binding principles are summarized in Figure 18.42 using as an example the small GTP-binding protein Ras (from *rat sarcoma*), a molecular switch that is inactive when bound to GDP, but activates a signal transduction pathway when bound to GTP. Ras specifically binds guanine nucleotides. Its affinity for GTP and GDP is in the picomolar range, while that for GMP is micromolar. Mg^{2+} is a required metal ion for high affinity: in its absence, GTP and GDP bind about 500-fold less tightly. The selectivity for the guanine base in GTP is established by two hydrogen bonds from a conserved Asp side chain. Any other nucleobase is rejected due to hydrogen bonding incompatibility. Both faces of the guanine base are in van der Waals contact, one with a hydrophobic Phe side chain and the other with the methylene groups of a Lys. These contacts contribute to the overall affinity but cannot explain the reduced affinity for GMP compared to GDP and GTP. The charges on the β-, and γ-phosphates are neutralized by the positive charges of a Lys side chain and a Mg^{2+} ion. Further charge dispersion is achieved by a network of hydrogen bonds from NH-groups of the phosphate-binding, or *P-loop*, and the positive end of a helix dipole immediately following the P-loop. The extensive hydrogen bond network connecting phosphates, water, protein, and the Mg^{2+} ion contributes to the high affinity of GTP (K_d = 10 pM) and GDP (K_d = 20 pM). The network is partially destroyed upon hydrolysis of GTP to GDP, but still strong enough so that GDP binds only two-fold weaker than GTP. Removal of the β-phosphate, *i.e.* hydrolysis of GDP to GMP, leads to loss of the Mg^{2+} ion, which drastically reduces the hydrogen bond network such that GMP has only weak affinity for Ras. Similarly, removal of the Mg^{2+} ion, for example by a metal-chelating agent, leads to breakdown of the network and the observed 500-fold loss of affinity for GTP and GDP.

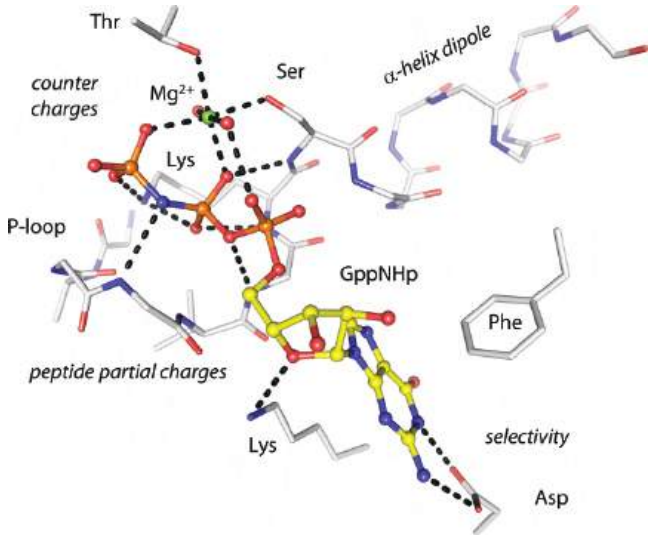


FIGURE 18.42: GTP-binding in Ras. A non-hydrolyzable GTP-analog, GppNHp, is bound to the human Ras p21 protein (PDB-ID 5p21). An Asp side chain establishes selectivity for guanine nucleotides by two hydrogen bonds (black dashed lines). The guanine base is sandwiched by non-specific van der Waals contacts with Phe and Lys side chains that contribute somewhat to affinity. The charges on the α- and β-phosphates are neutralized by a Mg^{2+} ion (green sphere) and a nearby Lys side chain. Further charge dissipation is accomplished by the P-loop and an α-helix, both of which orient their NH-groups to the phosphates. Only the backbone of the α-helix that orients its dipole toward the phosphates is shown. The extensive hydrogen bond network between the phosphates, water, protein, and the Mg^{2+} ion explains the high affinity for GTP and GDP.

Binding sites for ligands range from shallow depressions on the surface of the protein to internal pockets buried deep within the protein. Larger ligands such as oligosaccharides and sizeable enzyme substrates tend to bind in surface depressions or grooves. These shallow binding surfaces facilitate association and dissociation without expenditure of energy for large conformational changes by the protein. In contrast, permanently bound prosthetic groups such as heme, chlorophyll, pyridoxal phosphate (Figure 21.13 and Figure 21.38), but also structural metal ions and iron-sulfur clusters, are bound in deep pockets. Removal of such ligands often destabilizes the entire protein, and the ligand has to be considered as an integral part of the protein structure.

Binding sites are either pre-formed in the protein in the absence of ligand or formed upon ligand binding. The *lock-and-key* model of ligand binding implies a pre-formed binding site on the protein with little flexibility, which consequently has high selectivity for a certain ligand. This model describes an extreme case for a binding process because it posits that no conformational changes of the protein are required. An example for an enzyme operating predominantly by the lock-and-key mechanism is ribonuclease, with its rigid, pre-formed active site for RNA binding.

Conformational changes, both by the protein and by the ligand, are required in the majority of binding events. Two other models, *induced fit* and *conformational selection*, describe ligand binding associated with conformational changes of the protein. In the induced-fit model, a ligand weakly binds to a sub-optimal protein conformation and induces a conformational change in the protein to best accommodate the ligand. The protein is viewed as rather static, switching only between a few conformations. Hexokinase is an example for this binding mechanism: glucose binds and induces a conformational change in hexokinase that activates its phosphorylation activity. In contrast to the induced-fit model, the conformational-selection model assumes that proteins are rather malleable and change conformations rapidly between many states of similar energy. One of these states is highly complementary to the ligand and hence is selected for binding. Binding of the ligand removes this state from the conformational equilibrium. According to the principle of Le Chatelier (Section 2.1.2), the conformational equilibrium will keep re-populating this state. The continuous removal (selection) of the binding-competent state by ligand binding, will eventually "pull" the equilibrium

entirely to this state. The enzyme adenylate kinase, which catalyzes the reaction AMP + ATP ⇄ 2 ADP, uses this mechanism. The induced-fit and conformational-selection models are two arms of a thermodynamic cycle (Figure 18.43; Section 1.3.5).

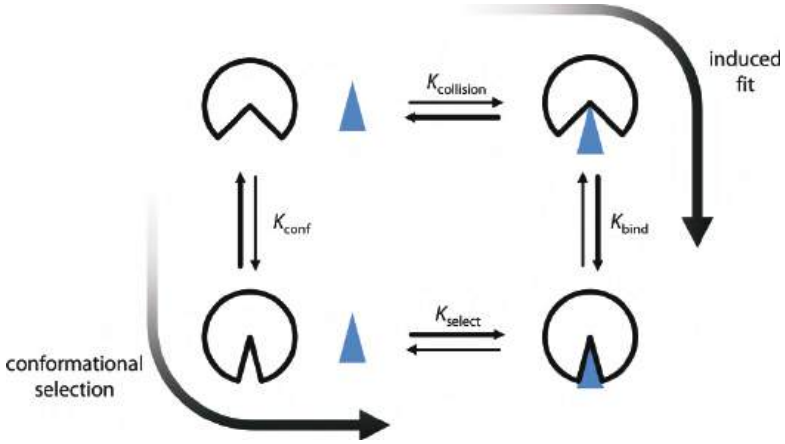


FIGURE 18.43: Induced fit *versus* conformational selection. In the induced-fit model, a ligand (blue triangle) forms a weak complex with a protein conformation that does not match the ligand's shape (binding constant $K_{collision} \ll 1\ M^{-1}$). Closure of the protein around the ligand leads to the high-affinity complex. Although energy is needed to induce the binding-competent conformation of the protein, this energy is overcompensated by the energy released by ligand interactions, so $K_{bind} \gg 1$. In the alternative conformational-selection model, the protein samples many conformations at low population ($K_{conf} \ll 1$). The ligand selects the best-fitting shape to form the high-affinity complex ($K_{select} \gg 1\ M^{-1}$). The thermodynamic cycle posits that $K_{collision} \cdot K_{bind} = K_{conf} \cdot K_{select}$.

While some ligands exclusively follow one of these models, the actual mechanism of protein-ligand complex formation is often a mixture of these extremes. Binding is always associated with conformational changes, always in the protein and very often also in the ligand. It is the extent of the conformational changes that varies. Proteins with an essentially pre-formed ligand binding site undergo only small conformational changes on the level of side chain rotamer changes. A certain amount of flexibility, *i.e.* conformational entropy, is required for fast association of the protein with the ligand. More sizeable surface loop movements exceeding 1 nm are necessary for binding of large ligands, *e.g.* co-substrates such as SAM or NADH, to their enzymes. Even larger conformational changes include rigid body domain rearrangements, as in hexokinase, and protein folding around a ligand. Folding of an unstructured region around a ligand into a defined conformation is an extreme example, often observed with transcription factors binding to DNA (Section 19.5.4.1). Rigid body domain movements are triggered by ligands that bind at protein interfaces in oligomers. Ligand binding to *allosteric* sites, *i.e.* sites remote from the active site of an enzyme can also induce rigid body domain movements, which in turn change the catalytic efficiency of enzymes (*e.g.* by non-competitive inhibition; Section 13.6.3).

In addition to protein flexibility, the entropy of the ligand also plays an important role for the overall binding energy. The more flexible the ligand, the higher the entropic cost for binding in a defined conformation. One aim in structure-based drug design is to pre-organize the conformation of a ligand by suitable substituents or intramolecular hydrogen bonds such that the entropy loss upon binding is minimized and the overall affinity is thus increased.

The affinity of ligands for their protein can be very high with K_d values ranging from the nanomolar to the sub-picomolar range. For example, the *biotin-streptavidin* interaction is among the strongest known in nature (K_d ca. 10^{-15} M) and this interaction is frequently used to select biotinylated macromolecules from complex samples. By contrast, metabolites (substrates and products in ES and EP complexes) usually have very low affinity for their enzymes, indicated by K_M values in the high micromolar or sometimes millimolar range. Enzyme catalysis necessitates tight binding to the

transition state (Section 16.2), but low to moderate affinities for both the substrates and the products to ensure high substrate turnover. The low affinity for substrates is often due to substrate or enzyme distortion required for catalysis. Part of the binding energy is converted into potential energy in the ES complex, often by changes in torsion angles, which can then be released during catalysis or product release. An important aspect to mention here is that ligand binding in the cell does not occur under ideal conditions. The millimolar concentrations of proteins in the cell increase the viscosity of the cytoplasm, limiting diffusion and rotation of macromolecules compared to an ideal solution. The ensuing proximity effects, *i.e.* hindered dissociation of a complex due to molecular crowding, can considerably increase the affinity of a ligand for a protein. The effect is particularly relevant for the two-dimensional diffusion of molecules in membranes.

18.4 MEMBRANE PROTEINS AND THEIR LIPID ENVIRONMENT

Most of the Sections in this Chapter involved soluble proteins in aqueous solution. However, about 20–30% of the human proteome are membrane proteins, which are embedded in lipid bilayers, or *membranes*. Membranes have very different properties compared to water, and membrane proteins have different structures compared to water-soluble proteins. Before we come to the structure and properties of membrane proteins (Sections 18.4.4 and 18.4.5), we will first introduce the membrane constituents, the lipids (Sections 18.4.1, 18.4.2), and the driving forces for membrane formation (Section 18.4.3).

18.4.1 Biological roles of lipids and membranes

The main biological functions of lipids are energy storage, signal transduction, and the formation of membranes. *Triglycerides* or *glycerolipids* (Table 18.10) are highly reduced molecules with a high energy density, making them a very compact source of metabolic energy. Lipids and related molecules also act as hormones and second messengers in signal transduction processes. Examples include the sterol hormones estrogen, testosterone, and cortisol, which bind to nuclear hormone receptors and regulate metabolic processes and the blood pressure by activation of gene transcription. Diacylglycerol and phosphatidylinositol are second messengers generated after activation of G-protein coupled receptors (GPCRs) and are involved in the activation of protein kinase C and Ca^{2+}-mediated signaling, respectively. Importantly, most lipids tend to self-associate into various structures, of which bilayers (membranes) are the most relevant in biology (Section 18.4.3).

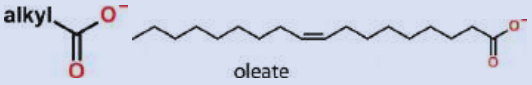
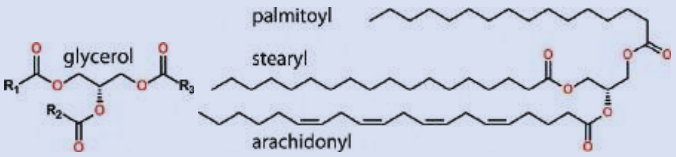
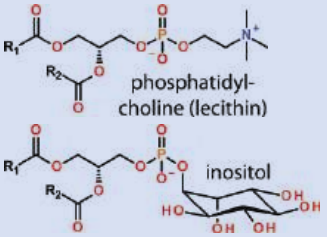
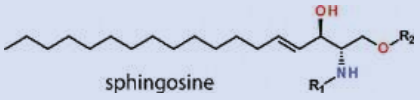
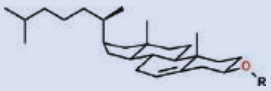
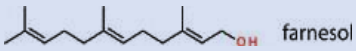
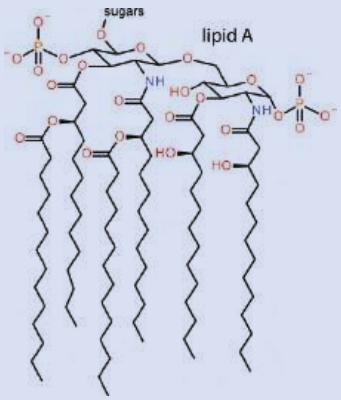
Membranes are not only the solvent for membrane proteins but also provide a barrier for electrical currents and molecules. These properties are exploited in living organisms to separate whole cells from the exterior environment. In addition, membranes subdivide eukaryotic cells into membrane-enclosed organelles surrounded by cytoplasm, including the nucleus, mitochondria, the endoplasmic reticulum (ER), the Golgi system, endosomes, lysosomes, and thylakoids in plants. Compartmentalization by membranes allows separation and thus regulation of biochemical processes. The ER and Golgi system play important roles in the post-translational modification and distribution of proteins (glycosylation and disulfide bond formation, Section 18.1.4).

The compartments separated by membranes become almost closed systems (Chapter 1.1): some exchange of heat but very limited exchange of matter occurs with the surroundings. While small hydrophobic molecules may passively diffuse through membranes, ions and larger molecules cannot pass. The impermeability of membranes for ions enables the storage of energy in the form of proton and other ion gradients (Section 4.2). The impermeability for protons, for example, is the prerequisite for the establishment of proton gradients (see Section 4.2) during oxidative phosphorylation in mitochondria and the photosynthesis in thylakoids. Another important proton gradient is maintained across lysosomal membranes, which establishes an internal pH $\approx$ 5 that facilitates hydrolysis of proteins taken up by endocytosis. Flow of matter across membranes is regulated by special membrane proteins (Section 18.4.4) known as channels and transporters (see Section 4.3). Signal-transducers such as GPCRs transport information across membranes by way of conformational changes.

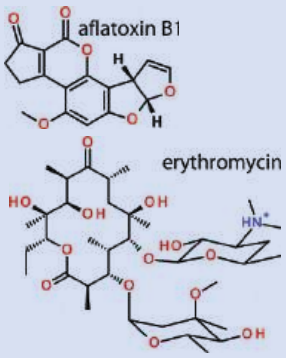
18.4.2 Types of lipids

Very generally, lipids are amphiphilic molecules with a hydrophilic *head group* at one end and a hydrophobic moiety at the other end. The two parts can vary widely, and more than 48000 different lipids have been reported in the Lipid Maps Structure Database (lipidmaps.org). Parameters that are varied in the hydrophobic part include chemical properties, size, length, as well as number, position, and stereochemistry of double bonds. The head groups are often charged, and can vary dramatically in size, from a hydroxyl group to a phosphorylated oligosaccharide. Based on the chemical classes of hydrophobic and head groups, lipids are classified into eight principal types: fatty acids, glycerolipids, glycero-phospholipids, sphingolipids, sterol lipids, prenol lipids, saccharolipids, and polyketides (Table 18.10). The major component of lipids in mammalian membranes are glycero-phospholipids, or just phospholipids (70–90%), followed by sterols, mainly cholesterol (10–20%), and sphingomyelin (5–10%).

TABLE 18.10:
Types of lipids according to the LIPID MAPS consortium.

type	structure, example, remarks
fatty acids	
glycerolipids	
glycero-phospholipids	 R_1, R_2 can be any fatty acid. Instead of choline the phosphoester can also be made with ethanolamine, inositol, and serine, all of which are found in membranes. In phosphatidylinositol phosphates (PIPs) the inositol can be phosphorylated to various extents. Note that inositol is not a carbohydrate. Glycosylation of inositol and acylation of the sugars with fatty acids results in glycophosphatidylinositol (GPI), a large family of lipids acting as membrane protein anchors.
sphingolipids	 R_1 is a fatty acid. If R_2 is H, the lipid is called ceramide. If R_2 = phosphatidylcholine, it is a sphingomyelin. For R_2 = glucose or a derivative, we have glycosphingolipids.
sterol lipids	 R = H: cholesterol R can also be a fatty acid or a sugar
prenol lipids	 Two isoprenyl moieties: geraniol also: retinol (vitamin A), β-carotene (Figure 21.13)
saccharolipids	 In Saccharolipids, the glycerol in glycerolipids or phospholipids is replaced by a sugar which forms esters and amides with up to seven fatty acids. Lipid A is part of the lipopolysaccharides in the membrane of Gram-negative bacteria.

polyketides



These are secondary metabolites synthesized by polymerization of acetyl and propionyl subunits. They exhibit extraordinary structural diversity: many are cyclic molecules that can be further glycosylated and methylated. Many are strongly toxic, or have antibiotic, anti-cancer, or anti-parasitic activities. They do not form membranes.

Detergents, also known as *surfactants* or *surface-active agents*, are synthetic amphiphilic molecules with similar biophysical properties to lipids. They are important tools in the solubilization of membrane proteins (Section 18.4.4) and for biophysical studies (Section 18.4.5). The common property of all amphiphiles is to reduce the surface tension at a polar/hydrophobic interface, *e.g.* water/air or water/oil.

18.4.3 Super-structures formed by amphiphiles

When mixed in low concentrations with water, amphiphiles are solvated: water molecules become ordered around the hydrophobic part in the form of a clathrate (a quite rigid, hydrogen-bonded network), which is an entropically unfavorable but enthalpically (slightly) favorable state. At higher concentrations, the monomeric amphiphile with its surrounding clathrate is no longer the most stable state but the amphiphiles start to interact with each other. The main driving forces for the assembly of amphiphiles into supra-molecular structures are the hydrophobic effect (Section 18.3.6.3) and, in charged amphiphiles, electrostatic repulsion of the polar head groups. Ordered water molecules are expelled from the hydrophobic surfaces of amphiphiles, increasing the entropy of water. The amphiphiles assemble into monolayers, micelles, or bilayers (membranes) by formation of many enthalpically favorable van der Waals interactions between the hydrophobic parts.

Amphiphiles partition into monolayers at the water/air interface, orienting their hydrophilic head groups toward the water and the hydrophobic parts toward the non-polar air (Figure 18.44). If the concentration of amphiphiles is increased above a certain threshold, known as the *critical micelle concentration* (*cmc*), the monolayers aggregate into different supra-molecular structures such as micelles, bilayers, tubes, and vesicles (Figure 18.44). All of these assemblies expose minimal hydrophobic surface to water. Micelles are densely packed, unilamellar, mostly spherical objects of 2–20 nm diameter and masses of typically less than 100 kDa that on average contain between 50 and 100 amphiphile molecules. In aqueous solution, the hydrophobic parts interact with each other in the center of the micelle and the hydrophilic head groups solubilize the whole micelle in water. Non-spherical micelles exist and in fact may strongly extend in one direction, forming tube-like shapes. An alternative to micelles is the formation of bilayers. The self-association of lipids into bilayers is key to the formation of biological membranes. We will use the terms *bilayer* and *membrane* interchangeably in the following. Bilayers are in principle infinite lateral aggregates of amphiphiles that expose only the hydrophobic rims to the polar solvent. This still unstable state can be avoided by spontaneous curving of the membrane into hollow spheres that no longer expose hydrophobic surface area to water. In biology, these spheres are known as vesicles. Their artificial counterparts, the *liposomes*, find medical application (Box 18.5).

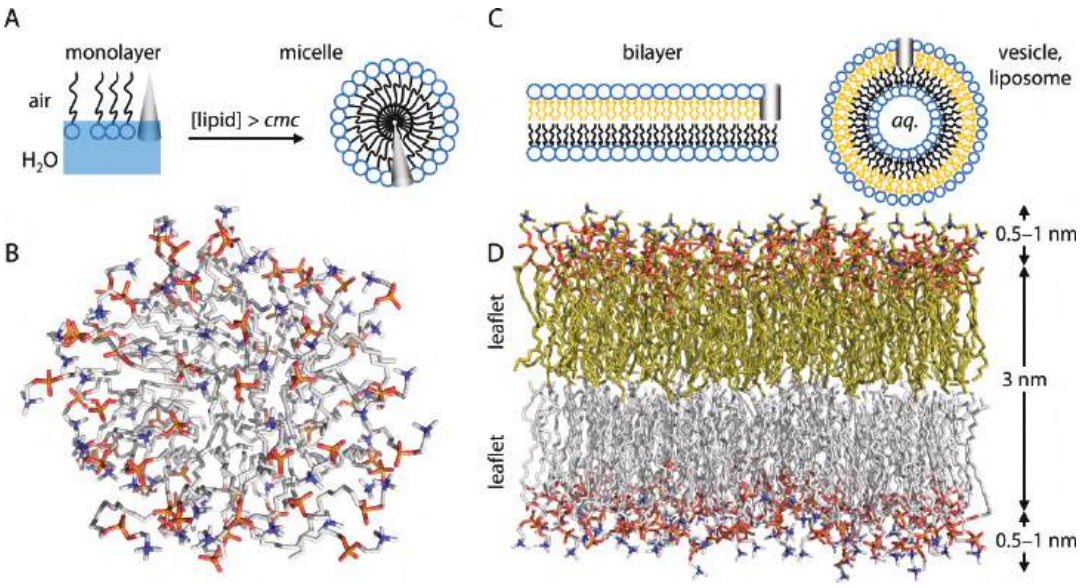


FIGURE 18.44: Assembly of lipids and detergents into monolayers, micelles, and bilayers. A: At a water/air or water/oil interface, amphiphiles first partition into a monolayer. If the shape of the amphiphile is conical and the concentration is larger than the *cmc*, unilamellar micelles are formed. B: This calculated micelle assembly consists of 65 molecules of *n*-dodecylphosphocholine (DPC), a zwitterionic lipid with only a single alkyl chain. C: Amphiphiles with larger hydrophobic chains have a more cylindrical shape and organize into membranes that bury the hydrophobic parts between two layers. The hydrophilic head groups are exposed on both sides toward the aqueous phases. Vesicles form when the bilayers curve into water-filled hollow spheres. Liposomes are essentially very small vesicles that can enable cellular entry of nucleic acids for gene therapy and vaccination. D: The simulated gel-like assembly of the membrane is colored yellow and gray to highlight the individual leaflets.

Exactly which type of assembly is formed by lipids depends not only on their chemical nature and concentration, but also on the temperature and, for charged amphiphiles, on the pH and the osmolarity of the solvent. The *detergent packing parameter P* of an amphiphile was defined to help predict the preferred shape of the assembly:

$$P = \frac{V}{A \cdot L} \qquad \text{eq. 18.3}$$

V and L are the volume and length, respectively, of the hydrophobic part of the amphiphile, and A is the cross-sectional area of the polar head group. P therefore depends on the molecular geometry of the amphiphile. Amphiphiles with $P < ⅓$, such as lipids with only a single fatty acid chain, are often cone-shaped molecules with wide hydrophilic head groups and slender hydrophobic tails. Such amphiphiles tend to form densely packed spherical micelles in aqueous environments. The molecular shape of amphiphiles with $⅓ < P < ½$ is more like a truncated wedge or a cone with its tip cut off. Their preferred aggregates are cylindrical micelles and tubes. Very hydrophobic amphiphiles with $P > ½$, such as glycerolipids, have about the same width along the whole molecule, leading to cylindrical molecular shapes that assemble into lamellar bilayers (membranes). The molecular shapes of amphiphiles influence their effects on membrane proteins solubilized in them. Molecular dynamics simulations on the conformational flexibility of amphiphiles can estimate the P value and hence predict their aggregation behavior. Amphiphiles assemble in a highly cooperative process.

18.4.3.1 Micelle formation

The assembly of monomeric amphiphiles S to micelles M_n is an interesting phenomenon because it follows a two-state process with no accumulation of intermediates:

$$n\,S \overset{K}{\rightleftharpoons} M_n$$

scheme 18.1

The *aggregation number n* is the average number of amphiphiles in the micelle, often in the range 50–100. Scheme 18.1 represents the equilibrium for an uncharged amphiphile. For charged amphiphiles, the expression becomes more complex because the salt concentration also needs to be taken into account. The strong cooperativity of micelle formation is reflected by the law of mass action, which contains the aggregation number *n* in the exponent of the monomer concentration [*S*]:

$$K = \frac{[M_n]}{[S]^n} \qquad \text{eq. 18.4}$$

The aggregation number *n* is a measure of the number of cooperative units. Micelle formation can be described by the all-or-nothing model (Section 3.3.1). For large *n*, the concentration of monomeric amphiphile becomes very small, and hence virtually impossible to measure, and the equilibrium constant *K* cannot be calculated. Useful proxies for measurement of micelle formation are heat, conductivity, turbidity, and refractive index, which all change strongly as a function of the known total concentration of amphiphile $[S]_{tot}$ when micelles form. $[S]_{tot}$ is given by

$$[S]_{tot} = [S] + n \cdot [M_n] \qquad \text{eq. 18.5}$$

Together with the equilibrium constant *K* defined in eq. 18.4, eq. 18.5 may be re-written as

$$[S]_{tot} = [S] + n \cdot K \cdot [S]^n \qquad \text{eq. 18.6}$$

in terms of monomeric amphiphile concentration [*S*] and

$$[S]_{tot} = \sqrt[n]{\frac{n \cdot [M_n]}{n \cdot K}} + n \cdot [M_n] \qquad \text{eq. 18.7}$$

in terms of micelle concentration $n \cdot [M_n]$. Plotting of [*S*] and $n \cdot [M_n]$ as a function of $[S]_{tot}$ displays a sharp bend near the *cmc* (Figure 18.45). The *cmc* is the concentration when aggregates of amphiphiles become soluble in the aqueous environment as micelles. It can in principle be estimated from the sharp bends described by eq. 18.6 for the monomeric species S and by eq. 18.7 for the micelle M_n. However, due to the high cooperativity of the monomer-to-micelle transition it is difficult to accurately measure [*S*] and $[M_n]$ near the *cmc*. Still, some useful approximations can be made regarding [*S*] and $[M_n]$. For $[S]_{tot} < cmc$, $[S] \approx [S]_{tot}$ and $[M_n] \approx 0$. Conversely, if $[S]_{tot} > cmc$, $[S] \approx cmc$ and $n \cdot [M_n] \approx [S]_{tot} - cmc$ (Figure 18.45).

The *cmc* and the aggregation number *n* are among the most important properties of an amphiphile. We will briefly outline ways to determine *cmc* and *n* from experimental data using the known total concentration of amphiphile $[S]_{tot}$. Once *cmc* and *n* are known, the equilibrium constant *K* may be calculated.

During a titration it is often much easier to determine the midpoint, or equivalence condition, than the individual concentrations of the reactants. In micelle formation, the special case $[S]_{equiv} = n \cdot [M_n]$ describes equivalence of the concentrations of free amphiphile [*S*] and the *n*-weighted micelles $n \cdot [M_n]$. Using this condition, eq. 18.5 may be re-written to

$$[S]_{tot} = 2 \cdot [S]_{equiv} \qquad \text{eq. 18.8}$$

This condition can be used to derive an expression from eq. 18.6 for $[S]_{tot}$ at equivalence that depends only on *n* and *K*, but not any more on the difficult to quantify individual reactants [*S*] or $[M_n]$:

$$[S]_{tot} = \frac{[S]_{tot}}{2} + n \cdot K \cdot \left(\frac{[S]_{tot}}{2}\right)^n \qquad \text{eq. 18.9}$$

Solving for $[S]_{tot}$ gives the total concentration at equivalence, $[S]_{tot,equiv}$:

$$[S]_{tot,equiv} = 2 \cdot \sqrt[1-n]{n \cdot K} \approx 2 \cdot \text{cmc} \qquad \text{eq. 18.10}$$

For large aggregation numbers $[S]_{tot,equiv}$ is about twice the *cmc*. If micelle formation is followed by *e.g.* isothermal titration calorimetry (Section 30.1), the experimental data allow extraction of $[S]_{tot,equiv}$ from the inflection point of the titration and hence offer a way to estimate *n*, *cmc*, and *K* by fitting (Section 30.1.2).

An alternative way to identify *cmc* is to measure the change in [*S*], d[*S*], divided by the change in total concentration $[S]_{tot}$, $d[S]_{tot}$, as a function of the total concentration. The plot $d[S]/d[S]_{tot}$ as a function of $[S]_{tot}$ is a sharp transition with an inflection point at the *cmc* (Figure 18.45).

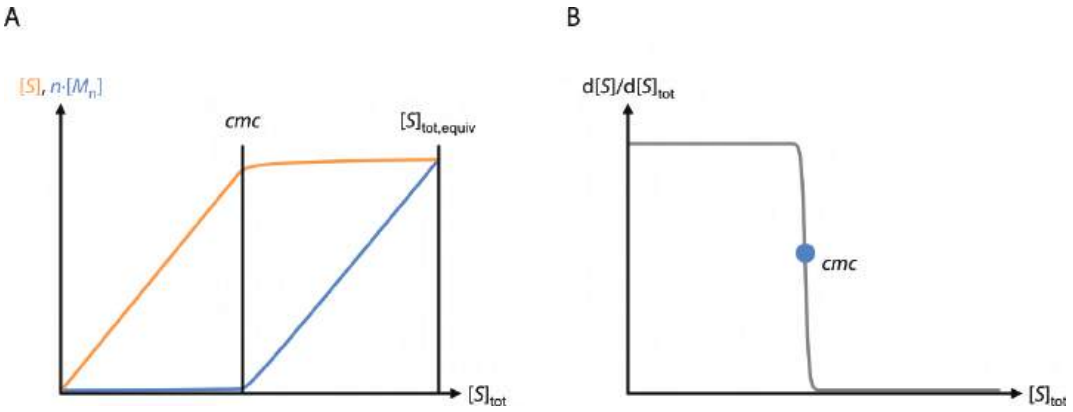


FIGURE 18.45: High cooperativity of micelle formation. A: Plot of eq. 18.6 for the monomeric species S (orange) and eq. 18.7 for the micelle M_n (blue). The sharp bend of both concentrations occurs near the *cmc*. The curves for the concentration of the two species intersect at $[S]_{tot} = 0$ and at $[S]_{tot,equiv} = 2 \cdot \sqrt[1-n]{n \cdot K}$, which for larger *n* is close to twice the *cmc*. B: The first derivative of the monomer concentration [S] with respect to the total concentration $[S]_{tot}$, $d[S]/d[S]_{tot}$, shows a very steep slope over a very narrow concentration range around the inflection point, which is at the *cmc*. The curve resembles an isothermal titration enthalpogram (Section 30.1).

The change in [*S*] over the change in total concentration $[S]_{tot}$, $d[S]/d[S]_{tot}$, is the first derivative of eq. 18.6. It must be calculated by implicit differentiation (Section 31.6.1) and is given by:

$$\frac{d[S]}{d[S]_{tot}} = \frac{[S]}{[S] + n^2 \cdot K \cdot [S]^n} \qquad \text{eq. 18.11}$$

While the concentration [S] of the monomer cannot be measured due to its low solubility and the high cooperativity of the transition, $d[S]/d[S]_{tot}$ is a quantity that may be measured directly for example by heat changes in ITC. From analytical calculus we know that the inflection point of eq. 18.11 is derived by setting its second derivative to zero. Since eq. 18.11 is the first derivative of eq. 18.6, the condition to obtain *cmc* from the inflection point of a titration is given by the third derivative of eq. 18.6 being zero:

$$\frac{d^3[S]}{d[S]_{tot}^3} = 0 \qquad \text{eq. 18.12}$$

When executing this derivative, finding the inflection point of eq. 18.11, and substituting that value into eq. 18.6, *cmc* can be calculated as the total concentration of amphiphiles, $[S]_{\mathrm{tot,cmc}}$, according to eq. 18.13:

$$[S]_{\mathrm{tot,cmc}} = \sqrt[1-n]{\frac{(2n-1)\cdot n^2 \cdot K}{n-2}} \cdot \frac{2n^2-2}{2n^2-n} \qquad \text{eq. 18.13}$$

This equation requires knowledge on n and K, which often is unavailable. However, the aggregation number n can be obtained by rearranging eq. 18.11 into a linearized form that does not depend on either *cmc* or K:

$$\frac{\mathrm{d}\ln\left(\frac{\mathrm{d}[S]_{\mathrm{tot}}}{\mathrm{d}[S]} - 1\right)}{\mathrm{d}\ln\left(\mathrm{d}[S]_{\mathrm{tot}}\right)} = \frac{n-1}{n} + \frac{(n-1)^2}{n} \cdot \frac{\mathrm{d}[S]}{\mathrm{d}[S]_{\mathrm{tot}}} \qquad \text{eq. 18.14}$$

Rearranging the measured $\mathrm{d}[S]/\mathrm{d}[S]_{\mathrm{tot}}$ data according to eq. 18.14 allows independent determination of the aggregation number n. Once n is known, either *cmc* or K need to be obtained to calculate the remaining parameter. If *cmc* is known from a fit to the data, the equilibrium constant K of micelle formation may be calculated from a rearrangement of eq. 18.13:

$$K = \left([S]_{\mathrm{tot,cmc}} \cdot \frac{2n^2-n}{2n^2-2}\right)^{1-n} \cdot \frac{n-2}{(2n-1)\cdot n^2} \qquad \text{eq. 18.15}$$

The free energy change ΔG_{m} per associating amphiphile molecule is then obtained by

$$\Delta G_{\mathrm{m}} = -\frac{RT}{n} \ln K \qquad \text{eq. 18.16}$$

Inserting eq. 18.15 into eq. 18.16 leads to an approximation for ΔG_{m} and large aggregation numbers n that is only dependent on the *cmc*.

$$\Delta G_{\mathrm{m}} \approx +RT \ln \mathrm{cmc} \qquad \text{eq. 18.17}$$

The *cmc* values of amphiphiles range from a few nanomolar for large hydrophobic chains to many millimolar for short chains. For example, in phosphatidylcholines that carry a single fatty acid, the *cmc* depends exponentially on the length of the fatty acid chain: five, six, and seven carbon atoms (pentanoic, hexanoic, and heptanoic acid) lead to *cmc* of 90 mM, 15 mM, and 1.4 mM, whereas phosphatidylcholines with 12, 14, and 16 carbon atom groups (lauryl, myristoyl, and stearyl) have *cmc* of 90 nM, 6 nM, and 0.5 nM, respectively. This shows that micelles are more easily formed by lipids with increasing chain length. As detergents are frequently used for the purification of membrane proteins (Figure 18.49), the *cmc* of the detergent needs to be known to avoid excess micelle formation. Table 18.11 collects the properties of a few amphiphiles frequently used in biochemistry.

TABLE 18.11:
Properties of selected amphiphiles relevant for micelle formation.

amphiphile	lipid	head group	charge	*n*	*cmc* (mM)	*cmc* (%)
CHAPS/CHAPSO	cholyl	ammonio-sulfonate	zwitter	10–11	8	0.49
β-OG	octyl[1)]	glucose	neutral	27–100	18–20	≈ 0.53
CYMAL-5	aliphatic[1)]	maltose	neutral	≈ 47	2.4–5	≈ 0.12
DDM	1 × lauryl	maltose	neutral	78–149	0.17	0.0087
DM	1 × decyl	maltose	neutral	≈ 69	1.8	0.087
SDS	1 × lauryl	sulfate	negative	62–101	8.2	0.23
Triton X–100	arom./aliph.[2)]	PEG	neutral	75–165	0.23	0.015
Tween 20	1 × lauroyl	sorbitan	neutral	n.a.	≈ 0.06	≈ 0.007
Brij 35	1 × lauryl	PEG	neutral	≈ 40	≈ 0.09	≈ 0.011
NP 40	nonylphenyl	PEG	neutral	100–155	0.05–0.3	≈ 0.02
LMNG	didecylpropane	ammonium- N-oxide	zwitter	≈ 76	1–2	≈ 0.023
LDAO	1 × lauryl	ammonium- N-oxide	zwitter	≈ 76	1–2	≈ 0.023
C8E4	aliphatic	hydroxyl	neutral	≈ 82	≈ 8	≈ 0.25
DPC	1 × lauryl	phosphocholine	zwitter	≈ 54	≈ 1.5	≈ 0.047
DOPC	2 × oleoyl	phosphocholine	zwitter	n.a.	0.05	n.a.

1) Cyclohexylpentyl.

2) (1,1,3,3-tetramethylbutyl)-phenyl. DDM: *n*-dodecyl-β-D-maltopyranoside. DM: *n*-decyl-β-D-maltopyranoside. LMNG: lauryl maltose neopentyl glycol. LDAO: lauryl dimethylamine-N-oxide. C8E4: tetraethylene glycol monooctyl ether. DPC: dodecyl phosphocholine. DOPC: dioleoyl phosphatidylcholine. Lauryl is the linear aliphatic C_{12} chain connected *via* an ether, lauroyl and oleoyl are the carbonic acids connected *via* an ester. The *cmc* is given for water at room temperature and will change with temperature. For ionic amphiphiles, *cmc* in addition depends on pH and osmolarity. n.a.: not available. Data taken from the Anatrace detergent properties booklet: https://anatrace.com/en/support-center/detergent-ooklet

18.4.3.2 Membrane formation and properties

Membranes are lipid bilayers where the hydrophobic parts of the lipids orient toward each other and form a non-polar central region that excludes water. The outward-facing surfaces of membranes are formed by the hydrophilic head groups of the lipids and are in contact with water. The two layers in membranes are also referred to as *leaflets* (Figure 18.44). In contrast to the comparatively small and mono-layered micelles, membrane bilayers can cover large areas with dimensions of up to a millimeter (Figure 18.44). Biological membranes have an overall thickness of about 4 nm, with the hydrophobic region covering about 3 nm and the hydrophilic parts adding 0.5–1 nm on each side. A layer of water molecules and ions is associated with the polar surface of membranes. While non-polar molecules such as gases and volatile anesthetics may diffuse freely across membranes, the amphiphilic properties of membranes makes them efficient diffusion barriers for ions. The negative charges on biological membranes determine their non-conductivity for anions, especially negatively charged metabolites (Box 18.4).

BOX 18.4: ELECTROSTATIC REPULSION OF METABOLITES BY MEMBRANES.

The non-permeability of membranes for negatively charged metabolites is used to keep initially neutral molecules within the cell by conferring negative charges onto them. This general principle in biology is exemplified by the phosphorylation of nutrients after their uptake into the cell. Glucose for example is transported by the glucose transporter from the extracellular milieu across the membrane into the cytoplasm and then rapidly converted into the negatively charged glucose-6-phosphate, a reaction catalyzed by hexokinase at the expense of ATP. Other kinases that serve to phosphorylate metabolites are fructokinase and galactokinase. In addition, many metabolites carry negative charges not only in the

(*Continued*)

form of phosphate groups but also as carboxylates. Examples include fatty acids, pyruvate, oxaloacetate, citrate, and many more. Intermediates of the biosynthesis of amino acids and nucleotides are also negatively charged. Electrostatic repulsion of these negative charges by membranes ensures that these molecules stay within their compartment and do not diffuse away. Electrostatic repulsion of therapeutic nucleic acids by the cell membrane is overcome by packaging them into liposomes (Box 18.5).

In contrast to anions, cations other than protons can overcome membrane barriers to some extent, especially if they are hydrophobic and have molecular masses < 500 Da. Hydrophobic cations, have amphiphilic character and are electrostatically attracted by the negatively charged biological membranes, which facilitates membrane penetration.

Two lipid bilayers can fuse in a process termed *membrane fusion*. Membrane fusion is very common *in vivo*. The reverse process, membrane fission, is also common. Examples of membrane fusion include exocytosis such as the fusion of neurotransmitter-containing vesicles with the plasma membrane at the synapse, and the shuttling of cargo (proteins, mostly) contained in vesicles between sub-cellular compartments. Membrane fission occurs as the last step during cell division and when it occurs within a cell it is part of endocytosis, the uptake of material from a cell's environment in the form of vesicles. Both membrane fusion and fission require close contact of the head groups, which is energetically unfavorable for the negatively charged biological membranes. Many intra-cellular membrane fusion events ("vesicle trafficking") are ATP-dependent and facilitated by membrane proteins of the SNARE family. Other membrane fusion events, such as cellular entry of enveloped viruses (*e.g.* vaccinia, HIV), are facilitated by SNAREs or by specific viral proteins. A well-known example is the role of hemagglutinin during vaccinia virus infection: once the virus is taken up by endocytosis, the pH of the endosome formed decreases, which triggers a conformational change in hemagglutinin that leads to the fusion of viral and endosomal membranes. Membrane fusion is also required for drug delivery using liposomes (Box 18.5). Here, artificial amphiphiles are added to overcome the electrostatic repulsion of the head groups to facilitate fusion.

BOX 18.5: FUSION OF LIPOSOMES WITH MEMBRANES AND THE USE OF LIPOSOMES IN DRUG DELIVERY.

Liposomes can bypass membranous barriers by membrane fusion. When liposomes fuse with the cell membrane, their content is emptied directly into the cytoplasm (Figure 18.46).

Liposomes filled with proteins have been used for enzyme replacement therapies in patients who, due to a mutation, lack the natural protein version. Examples are the lysosomal storage diseases Gaucher disease, where glucocerebrosidase has become functionally impaired, and Tay-Sachs disease, where β-hexosaminidase is mutated (Section 18.5.6). The most recent breakthrough when applying liposomes came with transporting therapeutic RNA into the cells. Membranes are usually impermeable to the negatively charged RNA, so liposomes are used to package the RNA and mask its charges. Once inside the cell after fusion, the RNA may serve different functions such as gene silencing (siRNA) or protein production (mRNA). mRNA is translated by the ribosome to the encoded protein with the purpose of treating cancer, genetic diseases, and infections. Viral proteins encoded by the mRNA, *e.g.* in the Pfizer/BioNTech and Moderna SARS-CoV-2 vaccines, are recognized by the immune system, which elicits an antibody response and confers immunity against future infections.

(*Continued*)

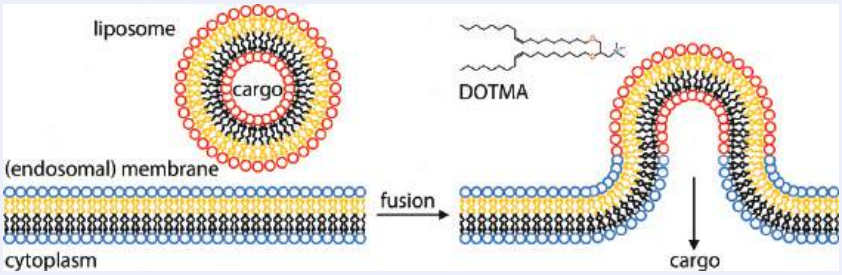


FIGURE 18.46: Spontaneous membrane fusion. The principle of liposome fusion with cellular membranes is the same as that between biological membranes. Liposomes filled with a cargo can either fuse directly with cell membranes or they are first taken up *via* endocytosis. Fusion with the plasma membrane delivers the content of the liposome into the cytoplasm of somatic cells (direct delivery). Endocytosis produces an endosome that engulfs the whole liposome with the problem that the cargo needs to reach the cytoplasm (indirect delivery). Endosomes mature into lysosomes of pH ≈ 5 and contain lipases that may prematurely hydrolyze the liposome membrane, at which stage the cargo would be lost. The cargo may still reach the cytoplasm by "endosomal escape" where the liposome inside the endosome fuses with the endosomal membrane. Both endosomal escape and direct fusion with the plasma membrane are facilitated by positively charged amphiphiles (red circles) that attract the negatively charged phospholipids. The structure of DOTMA, a cationic amphiphile often used for packaging SARS-CoV-2 vaccines in the form of mRNA, is shown to the right.

Liposomes are roughly classified into small, large, and giant with diameters of 0.02–0.1 µm, 0.1–1 µm, and 1–200 µm, respectively. The size depends on the mass of the cargo they need to transport. A mix of amphiphiles is used to optimize the size and biophysical properties of the liposome. For example, cationic amphiphiles such as DOTMA are used to stabilize negatively charged nucleic acid cargos. "Helper amphiphiles" such as cholesterol and phospholipids are mixed in to increase fusion efficiency and intracellular distribution.

18.4.3.3 Membrane dynamics and asymmetry

Membranes are highly dynamic structures. The lipids in a membrane diffuse laterally at a high velocity of about 2 µm s^{-1}. This means that a lipid in the membrane of a typical bacterium can travel from one end to the other in less than a second! Lateral diffusion velocities of suitably labeled lipids and membrane proteins can be measured by fluorescence methods such as FRAP (Section 21.5.5.3) or single-molecule tracking (Section 25.7.3). The assumption of free lateral diffusion of lipids and proteins in membranes is the underlying principle of the fluid mosaic model (Box 4.2). This assumption is not necessarily true, however. It has been found that the lateral diffusion coefficients of membrane proteins in eukaryotic membranes are often about tenfold smaller than expected from model membranes, possibly due to local crowding effects of proteins or interactions of membrane proteins with cytoplasmic proteins (Box 4.2). Also, the lipid composition in eukaryotic membranes is inhomogeneous. There are 10–200 nm sized patches ("rafts") enriched in cholesterol, sphingomyelin, and glycosphingolipids that change the local rigidity of the membrane. An important consequence of lipid rafts is restricted lateral diffusion of transmembrane proteins. The effect is known as corralled diffusion (Section 4.1).

For transverse diffusion or spontaneous *flip-flop* of a lipid from one layer to the other, the polar head group has to traverse the hydrophobic layer of the membrane, which is energetically disfavored. As a consequence, transverse diffusion is about nine orders of magnitude slower than lateral diffusion. Therefore, the contents of the outer and inner leaflets of the bilayer do not mix appreciably: biological membranes can maintain asymmetry of the leaflet contents. This is, for example, the case for erythrocyte membranes, where the outer leaflet has a high content of sphingomyelin and

phosphatidylcholine, whereas the inner leaflet is rich in phosphatidylserine. The asymmetry serves a biological purpose: rupture of a cell exposes phosphatidylserine from the inner leaflet to the extracellular medium, thus signaling the body that the cell is dying (apoptosis).

The lipid asymmetry of membranes is established and maintained by enzymes called flippases, floppases, and scramblases. Flippases transport lipids from the outer leaflet to the inner leaflet. Their activity is needed to restore lipid asymmetry after membrane disruption and repair. Floppases catalyze the reverse transport of lipids, from the inner to the outer leaflet. They are involved in membrane remodeling during endo- and exocytosis. Finally, scramblases catalyze the random movement of lipids between the leaflets. They can facilitate membrane fusion by homogenizing the lipid contents of the fusing membranes.

While discussing the principles of membrane formation, we touched on the important role of the lipid bilayer for the structure and stability of membrane-embedded proteins. In Section 18.4.4, we will get to know the different types of membrane proteins and the molecular principles underlying their interaction with membranes.

18.4.4 Properties and structure of membrane proteins

Membrane proteins are proteins that are permanently or at least most of the time associated with a membrane, which applies to about 20–30% of all proteins. Just as water is an important factor for the structure and stability of soluble proteins lipids are an integral part of the structure of membrane proteins, and have a decisive effect on their stability and flexibility. In essence, membrane proteins are amphipathic solutes dissolved in a membrane. The result is a two-dimensional solution with the central hydrophobic parts of the proteins placed in the hydrophobic solvent, flanked on both sides of the membrane by water-interacting hydrophilic parts. In general, polar groups in membrane proteins are either solvent-exposed, pointing away from the membrane, or they bind to the polar head groups of the lipids. Any polar groups or charges that are buried within the membrane must interact with other polar groups or other charges, similar to polar groups in the hydrophobic core of soluble proteins (Section 18.5.1) or within protein-protein interfaces (Section 18.3.6). Due to their amphiphilicity, membrane proteins display polarity: they are very rarely observed to flip within the membrane, as this requires energetically unfavorable transit of polar side chains through the lipid environment.

Membrane proteins are classified according to the extent to which they are associated with, or integrated into the membrane (Figure 18.47). *Monotopic* or *peripheral* membrane proteins interact with one leaflet of the lipid bilayer and may sometimes dissociate from it. *Integral* membrane proteins traverse the membrane and are thus permanently fixed to it. Depending on the number of crossings of the polypeptide chain through the lipid bilayer, integral membrane proteins are further subdivided into *bitopic* (single pass) or *polytopic* (multi-pass). Type I bitopic proteins have a cytoplasmic C-terminus and type II have a cytoplasmic N-terminus. Polytopic proteins have several passes of the protein structure through the membrane.

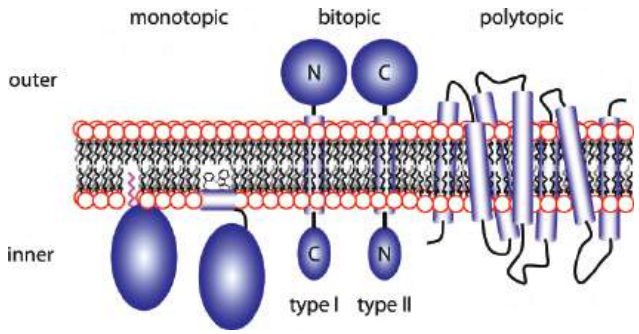


FIGURE 18.47: Classification of membrane proteins. Monotopic proteins are attached to a single leaflet of the membrane. They often serve to recruit soluble proteins from the cytoplasm to membranes, *e.g.* for vesicle trafficking. Bitopic proteins have a single transmembrane helix traversing both leaflets as their membrane anchor. The helix often connects two functional domains, *e.g.* an extracellular hormone receptor and an intracellular kinase domain, enabling signal transduction across the membrane upon hormone binding.

Most protein-membrane interactions in monotopic membrane proteins are hydrophobic, although some monotopic membrane proteins interact electrostatically with the polar head groups of the lipids. The hydrophobic interactions of monotopic membrane proteins with a single leaflet are mediated by amphipathic α-helices, hydrophobic loops, and lipids post-translationally attached to proteins. For example, the small GTP-binding protein Arf is anchored in the membrane by an amphipathic α-helix and an N-terminal myristoyl group. Ras and Rab proteins have C-terminal farnesyl and geranylgeranyl groups, respectively, covalently bound to a Cys side chain. Other monotopic proteins have a C-terminal glycophosphatidylinositol (GPI) anchor, a lipid with the same core structure as phosphatidylinositol phosphates (PIPs; Table 18.10) plus several saccharides, phosphates, and ethanolamine groups attached to the inositol. GPI anchors are attached to the C-terminus of enzymes such as choline esterases and phospholipases, but also to adhesion molecules, complement regulatory proteins, and receptors. The connection of monotopic proteins with the membrane is mostly irreversible, but with exceptions. The activity of Rab proteins, for example, is regulated *via* their transient removal from the membrane by binding to a specific protein known as Guanine nucleotide Dissociation Inhibitor (GDI).

In contrast to monotopic proteins, bitopic membrane proteins span both leaflets of the membrane by a single transmembrane α-helix. A continuous stretch of ca. 20 hydrophobic residues in the sequence of a protein is often diagnostic of the presence of a transmembrane helix. Generally, a hydrophobic peptide will form α-helices in a hydrophobic environment because in a helix most polar main-chain atoms become tied in hydrogen bonds and only few unfavorable contacts between polar peptide atoms and non-polar membrane atoms are left (Figure 18.22). Hydrogen bonding is thus a strong driving force for solvation of peptide groups and structure formation in a membrane environment. The side chains in the membrane-buried part of the helix are all hydrophobic to favorably interact with lipids. This is the principle behind the high stability of transmembrane helices. Bitopic membrane proteins have globular domains of varying sizes flanking the transmembrane helix that point into the different compartments on opposite sides of the membrane (interior and exterior). These parts point into different environments. Proteins in the cell membrane have cytoplasmic interior domains and extracellular exterior domains. Proteins embedded in the membrane of cellular organelles have their interior parts located in the lumen of the organelle (*e.g.* ER, lysosome, Golgi) and the exterior parts oriented toward the cytoplasm. Many cell surface receptors and protein kinases are bitopic membrane proteins and transduce extracellular signals across the membrane into the cytoplasm. For bitopic receptor kinases located in the plasma membrane, signal transduction generally involves binding of a ligand to the extracellular domain. Ligand binding induces a conformational change and dimerization of the receptor. The intracellular kinase domains brought together in this way will cross-phosphorylate each other and also other intracellular proteins, initiating a signaling cascade that is specific to the kinase activated.

The most complex type of membrane proteins are *polytopic* proteins, whose polypeptide chain passes several times through the membrane (Figure 18.48). Structure formation is driven by the hydrophobic effect, maximizing the number of hydrogen bonds within the membrane and leaving as few polar hydrogen bond donors and acceptors as possible in contact with lipids. These requirements limit membrane proteins to all-α or all-β (93% and 7% of all 9300 known structures to date, respectively). α/β structures cannot satisfy this condition as they have open β-sheets whose first and last strands expose polar peptide groups to the lipids (Section 18.3.2). Exposure of the unsatisfied hydrogen bond donors and acceptors in a buried β-sheet to the hydrophobic membrane interior would destabilize the entire protein. No integral membrane protein with α/β structure has been identified to date.

Structurally, all-α membrane proteins form α-helical bundles; all-β proteins form β-barrels. The hydrogen bonds in the membrane-embedded parts are maximized. These hydrogen bonds are stronger than in aqueous solution due to (1) the lower relative dielectric constant in membranes compared to water and (2) the absence of water competing for hydrogen bonding in membranes. In α-helical membrane proteins, the helices have hydrophobic sides that interact with the membrane, and more hydrophilic sides that enable interactions between helices. The entire protein then often

assembles further into oligomers. In β-barrels, hydrophobic side chains on the outside of the barrel mediate interactions with the membrane.

Important classes of polytopic membrane proteins include enzymes, transporters, pore-forming proteins such as (aqua)porins, ion channels, (GPCR) receptors, and proteins for cell adhesion. Porins, for example, are water-filled channels in the outer membrane of Gram-negative bacteria that allow passive diffusion of small (< 600 Da) molecules that would otherwise not be able to cross membranes. Depending on the channel diameter and electrostatic properties, even polar or charged molecules such as carbohydrates, ions, and amino acids can cross the outer membrane.

Members of the large class of GPCRs (ca. 800 are encoded by the human genome) bind to an extracellular ligand, *e.g.* a hormone or a neurotransmitter, which induces a conformational change that relays the signal across the cell membrane to intracellular signaling proteins. For example, the β_2-adrenergic receptor (β_2-AR) is activated by extracellular catechol-monoamine neurotransmitters such as adrenalin, and after a conformational change binds to the intracellular stimulatory G protein (G_s) (Figure 18.48). In the cytoplasm, this GPCR/G_s complex activates adenylyl cyclase, which in turn produces the second messenger cAMP for signal transduction.

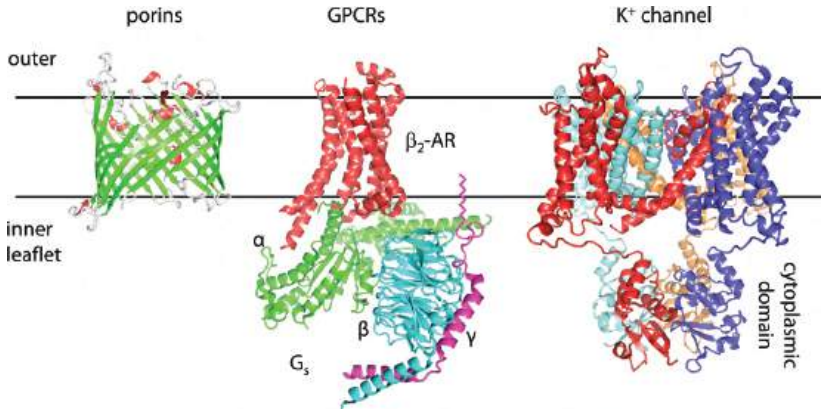


FIGURE 18.48: Examples of membrane protein structures. The top and bottom black lines delineate the extra- and intra-cellular space separated by the membrane. Bacterial porins usually form trimers in the membrane but for clarity only a monomer is shown (PDB-ID 1a0s). The β_2-adrenergic receptor (β_2-AR, red) is a eukaryotic G-protein coupled receptor that relays extracellular signals to the cytoplasm by docking to the heterotrimeric G-protein G_s (PDB-ID 3sn6). The receptor is a polytopic transmembrane protein and the monotopic heterotrimeric G-proteins are attached to the inner leaflet by a prenyl (farnesyl or geranyl-geranyl) chain at the C-terminus of the γ-subunit (magenta). The heterotrimer binds only to a certain conformation of the GPCR. The K^+ channel is a tetramer with a central polar path for the ions. Only the membrane-bound α_4-subunit is shown (PDB-ID 4jta). An intracellular domain of α/β fold is loosely attached to the all-α transmembrane part.

While GPCRs work as monomers that form transient hetero-oligomeric complexes, the K^+ channel is an example of a permanent functional homo-oligomer in a membrane. Four monomers arranged as a planar tetramer enclose a central ion-conducting channel. The diameter of the channel and specific ion-interacting side chains lining the inside of the channel account for exquisite ion selectivity.

β-barrels in membrane proteins can be entirely antiparallel (even number of strands) or have two parallel strands (the first and last in odd numbers of strands). Regardless of their topology, the face of the barrel that contacts the membrane is invariably made of hydrophobic side chains, suggesting that every other residue in a transmembrane strand should be hydrophobic. However, depending on the function of the protein, the inner face of the barrel can be either hydrophilic (as in porins) or hydrophobic, disrupting the pattern in the sequence. In contrast to transmembrane α-helices it is therefore not easy to glean a β-barrel membrane protein from its sequence. The task is easier for α-helices, although in case of polytopic membrane proteins, the helices may be kinked with respect to the membrane, so a larger number of hydrophobic residues is required for the helix to become buried in the membrane (Figure 18.48).

Polar loop regions connecting the β-strands of barrels or connecting the α-helices in helical bundles protrude from the membrane on both sides. These loop regions can be short simple β-turns (Section 18.3.3) or encompass whole domains that serve a biological function. Such domains may often be produced as soluble proteins, which greatly facilitates their study. In contrast, the study of entire membrane proteins is much more difficult as they usually need to be removed from their native, hydrophobic environment and solubilized in aqueous solution.

18.4.5 Solubilization of membrane proteins

We have seen in Section 18.4.4 that the membrane determines the structure and stability of membrane proteins. Functional studies of membrane proteins are therefore best done within their native membranes. This is possible with conductivity measurements on single ion channels using patch-clamp methods (Section 5.8), or with cellular assays that exist for many signal transduction processes. Most biophysical studies on membrane proteins require samples with higher concentration and purity. In the preparation of such samples, the membrane protein is removed from the biological membrane, exposing the hydrophobic rim that normally interacts with the hydrophobic membrane interior. These parts need to be solubilized, which is achieved by adding amphiphiles to the aqueous buffers that serve to mimic the lipid environment (Table 18.11).

The solubilization of membrane proteins is often done in the presence of non-natural, synthetic amphiphiles, the detergents. In contrast to many lipids, detergents have at least some solubility as monomers in water, which is a prerequisite to gently extract membrane proteins and stabilize them in solution. When mixed with polar and non-polar solvents, detergents behave similarly to lipids, and form monolayers, micelles, membranes, and vesicles. Although few commonly used detergents form membranes on their own, the aggregation number n and *cmc* are good estimates for how strongly their micelles interact with membranes. Detergents with small *cmc* and large n tend to display higher affinity for membranes, which is beneficial for membrane protein solubilization.

The actual detergent used depends on the measurements to be done. Finding the best amphiphile is mostly a matter of trial and error, aided by experience and statistics on previously successful preparations. SDS is useful for the purification of proteins under denaturing conditions, which can be an advantage if the purified protein can be refolded or if the native structure of the protein is not important for the experiment. Triton X–100 and NP 40 are also quite efficient in solubilization of membrane proteins but display high UV absorption, precluding many spectroscopic analyses. CHAPS, CHAPSO, β-OG, and C8E4, are examples of non-denaturing detergents that solubilize membrane proteins while retaining their structure. They share the advantage that they can be removed by dialysis later. DPC is commonly used for NMR applications while β-OG, DM, and DDM are quite successful when the goal is crystallization of the purified membrane protein (Box 24.1).

To date, no detergent can stabilize a membrane protein as well as the native membrane does, in other words, membrane proteins solubilized by detergents are always less stable than the same protein in native membranes. The reduced stability of membrane proteins in detergent-containing solutions has several reasons. Detergents incompletely shielding the hydrophobic surfaces of membrane proteins will not prevent unfolding and aggregation. Some amphiphiles induce conformational changes in membrane proteins that may result in reduced stability. Finally, membrane tension or "lateral pressure" as a result of the hydrophobic effect in membranes was found to contribute to membrane protein stability. When removed from their membranes, this pressure drops, which may decrease the stability of the membrane protein.

The two most successful general procedures in using detergents for membrane protein purification are *micellar extraction* and reconstitution in *nanodiscs*. In micellar extraction (Figure 18.49) the membrane protein is solubilized in an aqueous buffer containing an amphiphile at a concentration only slightly (1–2-fold) larger than its *cmc*, which ensures that both micelles and monomeric amphiphiles are present. The amphiphiles start to compete with the lipids of the natural membrane, and an "extended" micelle eventually encircles the hydrophobic part of the membrane protein. The arrangement of amphiphiles in such micelles is very different from that in bilayers, which may explain the inefficient stabilizing effect of amphiphiles using this method.

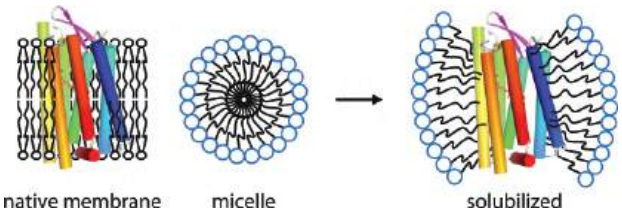


FIGURE 18.49: Micellar extraction of membrane proteins from their native membranes. The mono-layered micelle eventually forms a ring-shaped structure around the hydrophobic part of the membrane protein that is normally buried in the bilayer.

A more recent development is the reconstitution of membrane proteins in nanodiscs (Figure 18.50). Nanodiscs have revolutionized membrane protein research because membrane proteins in nanodiscs may essentially be treated as soluble entities, without the usual problems of instability and unspecific aggregation. As the name implies, nanodiscs are disc-shaped aggregates of amphiphiles. Their diameters range between 10–20 nm, depending on the size of the membrane protein they envelop. In contrast to micelles they retain the bilayer structure of natural membranes, providing a much more natural environment for membrane proteins than micelles do. However, they in principle suffer from the same problem as any membrane fragment in that nanodiscs would expose a hydrophobic rim to the aqueous solvent, which would make the disc unstable. The problem is solved by an amphiphilic scaffold, which wraps around the hydrophobic rim and shields it from water. In practice, the scaffold is added to an amphiphile solution (phospholipids, phosphatidylcholine, or a mixture of amphiphiles) and converts the micelles present into nanodiscs.

Scaffolds can be either a protein or a synthetic polymer. Protein scaffolds are derived from human apolipoprotein A-I, a component of high-density lipoprotein particles in our plasma. Apolipoprotein A-I is a helical and amphiphilic protein. The hydrophobic side of the helices serve as the scaffold for amphiphiles to aggregate. The hydrophilic opposite side of the helices helps to dissolve the nanodisc in water. Nanodiscs typically have one or two copies of apolipoprotein A-I.

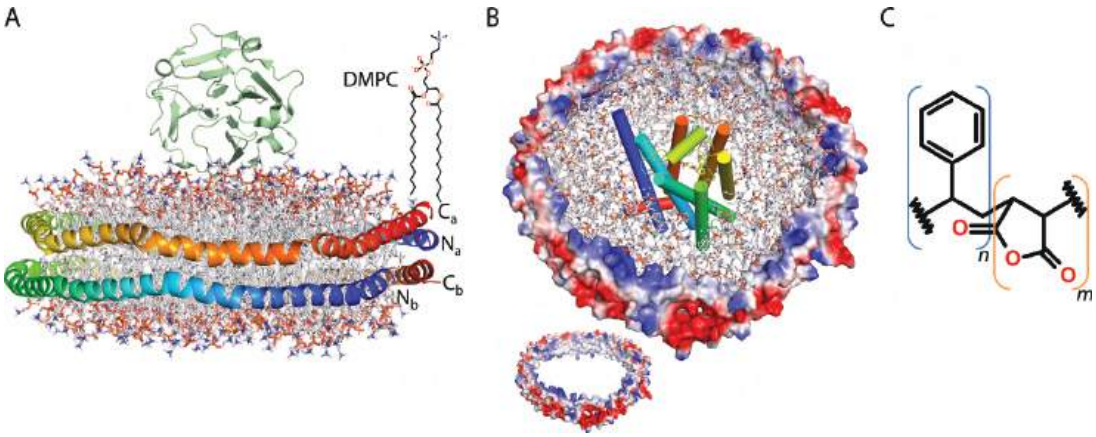


FIGURE 18.50: Nanodiscs for the stabilization and study of membrane proteins. A: Structure of a nanodisc (PDB-ID 6clz). The two layers, made of 118 DMPC molecules (1,2-dimyristoyl-sn-glycero-3-phospho-choline), are colored with gray myristoyl chains. The detergent-only part of the disc has a dia-meter of ca. 9 nm. The hydrophobic rim of the disc is covered by two copies of the amphiphilic protein human apolipoprotein A-I, which wraps around the disc as a long, bent α-helix. The two apolipoprotein monomers (a and b) are rainbow-colored from blue at the N-terminus (N_a and N_b) to red at the C-terminus (C_a and C_b), indicating that the dimer is anti-parallel. Here, a four-bladed β-propeller domain (light green) of human matrix metalloproteinase 14 anchors the protein protease by inserting a hydrophobic loop into the top layer of the nanodisc. B: Model of a transmembrane protein, the human β_2-adrenergic G protein-coupled receptor (PDB-ID 2rh1), inserted into a nanodisc. The electrostatic potential of apolipoprotein A-I is shown in red (negative), blue (positive), and gray (hydrophobic). The inset without detergent shows that the inner surface of apolipoprotein is hydrophobic while the outer surface solubilizing the nanodisc is polar. C: Styrene-maleic acid polymer as an alternative to apolipoprotein.

Amphiphilic styrene-maleic acid polymers are even more versatile as scaffolds. They can be synthesized in different lengths and with different ratios of the monomers styrene and maleic acid anhydride. Changing the length and component ratio of the scaffold allows fine-tuning of the nanodisc size and its overall solubility. The styrene assumes the hydrophobic part and the maleic acid anhydride, which is hydrolyzed after polymerization, confers negative charges on the nanodisc. In nanodiscs, membrane proteins can be analyzed by solid-phase methods such as TIRF microscopy (Section 25.4) and SPR (Section 29.3). Membrane protein preparations in nanodiscs have also proven useful for structural studies by cryo-EM (Section 26.5.2).

In Section 18.4, we have seen that the stabilization of soluble and membrane proteins follows the same principles and involves similar types of interactions. In both cases, the solvent plays an important role for stability. Until now, we have focused on the structures themselves. In Section 18.5, we will turn to the process that generates this structure, the folding of the polypeptide chain into a compact three-dimensional shape, and the underlying energetics and kinetics.

18.5 FOLDING AND STABILITY

The thermodynamic stability of proteins under physiological conditions is small: the free energy G_N for the native state is often just a few kJ mol^{-1} more negative than G_U for the unfolded state. The native state has many attractive interactions but low entropy, whereas the unfolded state has few attractive interactions but high entropy (see Box 1.11). Therefore, both G_N and G_U are large negative numbers. The net stability of the protein under specific conditions is the small difference $G_U - G_N$, which is typically around 20–40 kJ mol^{-1}. This corresponds to a small negative ΔG for folding of proteins, which is a result of entropy-enthalpy compensation (see Box 1.9). The enthalpic change during protein folding is dominated by differential interactions between the side chains in the folded state and of the side chains with water in the unfolded state. The entropic change is a balance between the reduced entropy of the folded state and the increased entropy of the water that is released upon folding. The folding equilibrium between native and unfolded states is easily shifted as a function of temperature and solvent composition. There are two possible strategies to destabilize a protein, either by destabilizing the native state or, maybe less intuitively, by stabilizing the unfolded state (Figure 18.51).

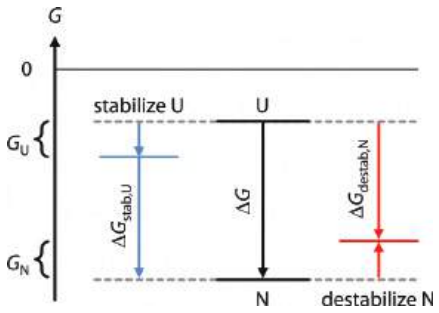


FIGURE 18.51: Dependence of protein stability on the energies of the folded and unfolded states. Black lines indicate the energy levels of a protein with the energy difference ΔG between the folded state N and the unfolded state U. The red lines show protein destabilization by destabilizing the folded state ($\Delta G_{destab,N}$). The blue lines show protein destabilization by stabilization of the unfolded state ($\Delta G_{stab,U}$). The protein is destabilized in both cases.

18.5.1 Driving forces for protein folding

We have introduced the hydrophobic effect, the tendency of non-polar molecules to self-associate in polar solvents and exclude water from the interface, in Section 18.3.6.3. The largest energetic contribution to the hydrophobic effect stems from the increase in entropy on release of water

molecules from hydrophobic surfaces. The hydrophobic effect is the driving force for the association of lipids to micelles, monolayers, in aqueous environments (Section 18.4). Similarly, it drives the clustering of hydrophobic side chains in proteins to exclude water and form the tightly packed hydrophobic core of the native state. Protein interiors are extreme environments, in the sense that they are more hydrophobic than protein-protein complex interfaces, and have less ionic interactions and less buried water molecules per surface area.

The transfer of non-polar side chains from the polar solvent water to the non-polar environment of the protein core has both favorable enthalpic and entropic contributions. The enthalpic contribution comes from van der Waals interactions between the side chains. The entropic contribution stems from the release of ordered water molecules on the surface of the hydrophobic side chains into bulk solvent. On the other hand, the release of water is associated with an energetic penalty: the strong hydrogen bonds between the ordered water molecules on hydrophobic side chains in the unfolded state (Figure 18.41) have to be broken, and the hydrogen bonds that re-form in bulk solvent are weak, which leads to an increase in enthalpy. The ordered water molecules around hydrophobic side chains in the unfolded state are the molecular reason for the high heat capacity of unfolded proteins compared to the folded state: folding is associated with a decrease in heat capacity C_p (Section 1.6.2, Figure 1.26).

A stable monomeric protein has few hydrophobic side chains on its surface. If present, they are mostly single and surrounded by hydrophilic side chains. If clusters of hydrophobic residues occur at the surface of a protein they are usually part of a protein-protein interaction site (Section 18.3.6).

While the burial of hydrophobic residues in the protein core is the main driving force for protein folding, hydrophilic side chains have less effect on protein stability. Hydrophilic amino acids decorate the solvent-exposed surface of soluble proteins where they form almost isoenergetic hydrogen bonds among themselves or with water. This is different for hydrogen bonds formed by main-chain and side-chain groups that are buried in the hydrophobic core of the protein: buried hydrogen bonds have a large energetic contribution to stability because there is no competition by hydrogen bond formation with solvent molecules and the interaction is stronger in environments of low dielectric constant (Section 17.2.2.3). Buried hydrogen bonds are also responsible for the strong *cooperativity* of protein folding (Section 18.5.4).

Although most of the water is removed from hydrophobic surfaces upon folding, water may still end up in a non-polar pocket of the protein interior. About 1% of the protein core is not completely packed with side chains. Sub-optimal packing of hydrophobic cores reduces van der Waals contacts between side chains and at the same time increases their entropy. Both effects are energetically unfavorable, and destabilize the protein. This is the reason why nature "abhors a vacuum" (*horror vacui*) in the protein interior. One of the strategies by which proteins in thermophilic organisms are stabilized compared to their mesophilic counterparts is exquisite packing of the hydrophobic core with few or no cavities. However, if a cavity is larger than the volume of a water molecule (> 30 $Å^3$ or 0.03 nm^3), water being the solvent will have to fill it. A single water molecule in a hydrophobic environment is frustrated because it cannot form any hydrogen bond. A minimum of three water molecules must be trapped to form two hydrogen bonds each, which sets a minimum volume for any solvent-filled cavity. In fact, hydrophobic cavities in proteins are either tiny and empty or large enough to host at least three water molecules. Over 90% of water molecules restrained in hydrophobic cavities are highly organized and form 3–4 hydrogen bonds per molecule.

Not only water as the solvent affects protein stability. Co-solutes can have stabilizing or destabilizing effects. The most prominent examples are salts, which can selectively stabilize the native state of a protein or the unfolded state, leading to a stabilization or destabilization of the protein, respectively. The Hofmeister series ranks the observed effect of salts on protein solubility and stability (Box 18.6).

BOX 18.6: THE HOFMEISTER SERIES AND THE EFFECT OF CO-SOLUTES ON PROTEIN STABILITY.

The effect of salts on protein solubility was tested ~1880, leading to the empirical *Hofmeister series* of cations and anions (Figure 18.52).

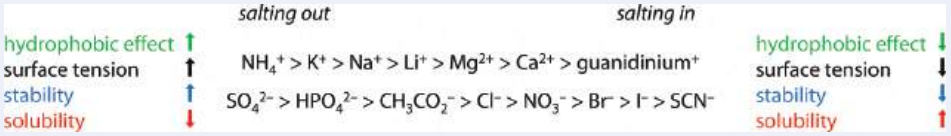


FIGURE 18.52: Effect of salts on protein solutions. The effects of salts on solubility correlate with surface tension and protein stability. The underlying biophysical cause of protein stabilization and destabilization is a change in hydrophobic effect.

Salts from the combinations of ions on the left of the series (such as ammonium sulfate) strongly reduce protein solubility (*salting out*) while those from the right-hand side of the series (such as guanidinium thiocyanate) strongly solubilize proteins (*salting in*). The effects on solubility are related to an increase or reduction of the hydrophobic effect. Salts like ammonium sulfate increase the hydrophobic effect and concomitantly decrease the solubility of proteins. The net result is the stabilization of the native state for the price of reduced protein solubility. Ammonium sulfate both stabilizes and precipitates proteins from aqueous solutions and therefore is often used for protein purification, concentration, crystallization, and long-term storage of precipitated proteins. By contrast, solubilizing salts from the right-hand side of the Hofmeister series decrease the hydrophobic effect and promote solubilization of non-polar surfaces in water. However, non-polar surfaces are a hallmark of unfolded proteins, so the net result is a destabilization of the protein by stabilization (solubilization) of the unfolded state. Ions of such *chaotropic* salts directly bind to the hydrophobic surfaces exposed in the unfolded state. Guanidinium ions, for instance (Figure 18.53), stack against peptide groups and aromatic side chains *via* cation-π interactions (Section 17.2.2.2). Urea (Figure 18.53) is a non-ionic chaotrope that forms hydrogen bonds with the peptide group, possibly competing with the native hydrogen bonding in proteins. Urea thus contributes to solvation of the unfolded state and induces protein unfolding. Basically any substance that efficiently competes for hydrogen bonding can act as a chaotrope. Some of these chemicals are used in the laboratory to remove proteins from solutions of nucleic acids (*e.g.* in phenol extraction and ethanol precipitation) or to solubilize denatured proteins (*e.g.* SDS for polyacrylamide gel electrophoresis). The alkyl chain of SDS (Figure 18.53) interacts with hydrophobic surfaces of aliphatic and aromatic amino acids, again solubilizing (and stabilizing) the unfolded state. While binding of SDS to hydrophobic side chains leads to destabilization of soluble proteins, membrane proteins are stabilized by small concentrations of detergents, including SDS, because they interact with hydrophobic surfaces present in the folded state (Section 18.4.4).

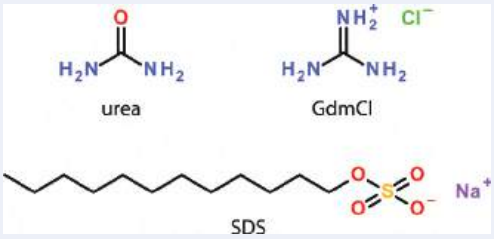


FIGURE 18.53: Structures of common chemical denaturants. Shown are urea, guanidinium chloride (GdmCl), and the detergent sodium dodecylsulfate (SDS).

In this Section, we have seen that the driving forces for protein folding are the same principles that stabilize protein structures. Now we can turn to the folding process itself: how is the compact, folded state reached?

18.5.2 First folding experiments and the Levinthal paradox

In 1956, Christian Anfinsen studied the dependence of RNase A activity as a function of reducing agent. RNase A is a small protein of 124 amino acids with four disulfide bonds that is deactivated by high concentrations of the denaturants urea and guanidinium chloride. Anfinsen unfolded RNase in urea and reduced the disulfide bonds with β-mercaptoethanol. When this inactive form was exposed to air, disulfide bonds re-formed in a random fashion. After removal of urea by dialysis, this *scrambled* RNase was almost inactive, retaining just 1% of its initial activity. However, upon addition of just a catalytic amount of reducing agent, the RNase regained its full activity through the process of *disulfide shuffling* (Figure 18.54).

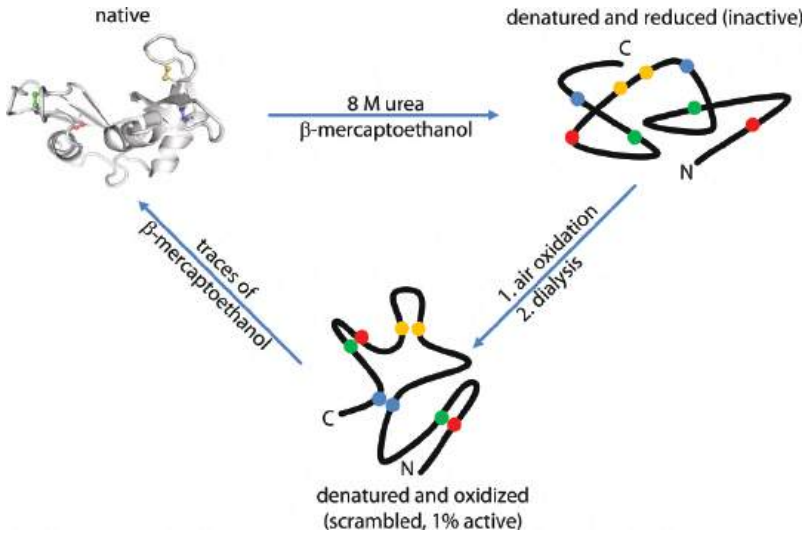


FIGURE 18.54: Anfinsen's experiment. Native RNase A is denatured and its four disulfide bonds are reduced with β-mercaptoethanol. Air oxidation in the presence of the denaturant allows the Cys residues to form random disulfide bonds. The denaturant is then removed by dialysis, which allows folding of this scrambled RNase A. Only 1% of the protein has the correct arrangement of disulfide bonds and displays activity, the other 99% remain unfolded and inactive. Adding a trace of β-mercaptoethanol catalyzes disulfide shuffling, and eventually all scrambled RNase A molecules will adopt their native state and become active.

It was apparent that the unfolded form of RNase A was trapped in an almost inactive state after air oxidation, indicating that incorrect disulfide bonds had been formed. Addition of only a catalytic amount of reducing agent was enough to completely reactivate RNase A, indicating that the driving force for reaching the native state comes from the protein. These impressive results show that the primary structure (the sequence) encodes all the information necessary for a protein to adopt its active three-dimensional structure.

If the information for the native structure is encoded in the polypeptide sequence, how does an unfolded polypeptide chain reach the native state? A simple Gedankenexperiment shows that it is impossible for the polypeptide chain to sample all possible conformations and find the native state by a random search. Let us assume that each amino acid residue in a polypeptide can sample ten different conformations. For N residues, we then obtain 10^N different conformations. The average rotation frequency around a chemical bond is ca. 10^{14} s^{-1}, so sampling of all conformations in just a 40-residue polypeptide would require $3 \cdot 10^{18}$ years (the universe is about $1.4 \cdot 10^{10}$ years old), which is clearly too long to support life (present on earth for only $3.5 \cdot 10^9$ years)! It is also much longer than

the folding times for typical proteins observed experimentally, which range from a few milliseconds to seconds. This discrepancy between the time scale of statistical sampling of the folding space and the true time needed for a protein to fold is known as the *Levinthal paradox*, formulated in 1969. It looks as if the primary structure not only encodes the end product of the folding process, but also the underlying kinetics. Exactly how this works remains a conundrum. However, if the protein structure is encoded in its sequence, then at least it must be possible to calculate the structure from the sequence. The *protein folding code* remained elusive for decades and *de novo* calculation of protein structures was impossible until recently (Section 20.3.2).

18.5.3 Energy landscapes for protein folding

The Levinthal paradox assumes that all conformations accessible to the polypeptide chain are sampled with equal probability until the native state, the conformation of lowest free energy, is found. The associated energy landscape (Section 17.3) is a flat surface with a single minimum, akin to the putting green in a golf course (Figure 18.55). This scenario is unrealistic not only for reasons of time, but also for energetic reasons: it assumes that all conformations of the polypeptide chain that do not correspond to the native state have equal energy. However, we know from the Ramachandran energy surface (see Figure 18.21) that certain conformations have lower energies than others. A more realistic energy landscape for folding is a *folding funnel*. In its idealized, smooth shape, all starting points at a perimeter have the same energy. This model correctly predicts that with each native-like contact that is established, ΔG becomes more and more negative, and the structure is closer to the native state. However, a smooth funnel model lacks local energy minima where folding intermediates can reside, and predicts that folding occurs without intermediates. This is not true; such intermediates have been observed experimentally, *e.g.* for proteins that contain disulfide bonds, *cis*-peptides, or several domains. Probably the best model to visualize protein folding is a *rugged funnel*, an extension of the smooth funnel. The energy landscape has local minima with activation barriers between them. Proteins that reach a local minimum with a high energy barrier will need time to overcome this barrier and to reach the native state. Local minima therefore lock the molecule in a *kinetic trap*, which has been evidenced experimentally for some proteins. In the rugged funnel model, there are several alternative pathways toward the native state: the folding of a protein may proceed without detectable intermediates or through one or more intermediates. It is this variability of possible folding scenarios that makes the rugged funnel model the most realistic energy landscape of protein folding.

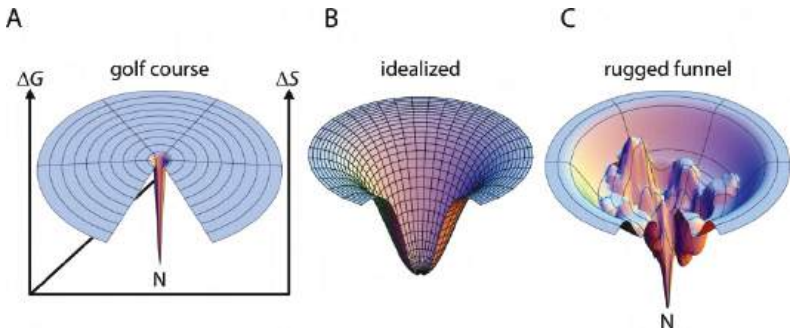


FIGURE 18.55: Energy landscapes of protein folding. A: In the "golf course putting green" landscape from Levinthal's paradox, all folding intermediates have equal energy, an impossible scenario. B: The smooth funnel landscape is more realistic because folding can start from any point on the energy surface and should move in the correct direction. C: The rugged energy landscape includes local minima for folding intermediates that can act as kinetic traps on the way to the folded state. Panels A and C reprinted from Dill & Chan (1997), Nat. Struct. Biol. 4(1): 10–19, with permission.

Although the funnel analogy is useful for visualizing protein folding and its energetics, the funnels must not be taken literally: folding funnels give the wrong impression that the shape of the energy landscape guides the protein toward the folded state. In fact, some functional proteins known as *intrinsically disordered proteins* do not even have a defined native state and often lack any regular structure (Box 18.7).

BOX 18.7: INTRINSICALLY DISORDERED PROTEINS.

Until the 1990s it was assumed that a functional protein needs a defined three-dimensional structure. The few exceptions of functional but unfolded proteins were limited to small peptide hormones such as glucagon, which were thought to fold upon binding to their receptors. The unfolded state of proteins was mainly considered non-functional, populated only during synthesis and chaperone-assisted folding, and before degradation. A sequence database survey conducted in 1994 showed that about half of all eukaryotic proteins known at that time contain stretches of > 40 residues that are of low complexity, *i.e.* contain sequences that will likely not fold into any defined structure. Eukaryotic genomes were found to encode more proteins with intrinsic disorder than those from prokaryotes. Nowadays a few dozen proteins are known to be entirely unfolded or *intrinsically disordered* under physiological conditions. Compared to folded proteins, intrinsically disordered proteins have a higher hydrophobicity and lower net charge, and are therefore more aggregation-prone. They display little, if any, secondary structure, and usually do not adopt a globular shape. Some of them, for instance α-synuclein in nerve cells, adopt helical secondary structure when in contact with phospholipids, while others, including the zinc finger-containing transcriptional repressor RYBP, have an unfolded DNA-binding domain that adopts a regular structure upon engagement with its target sequence.

18.5.4 Mathematical description of the two-state model

We have seen in Section 18.5.3 that folding can occur with or without intermediates. The simplest conceptual model for protein folding is the two-state model that assumes a thermodynamic equilibrium between only two states, the folded or native state N and the unfolded state U, without intermediates (Figure 18.56). By native state, we refer to the functional form of the protein. For small proteins, the native state is often a single conformation, for larger proteins it can be a set of defined conformations. By contrast, the unfolded state is a large ensemble of conformations that exhibit only local interactions. These conformations rapidly interconvert and have similar energies. This simplification corresponds to a minimalistic energy landscape similar to the golf-course scenario in Section 18.5.3.

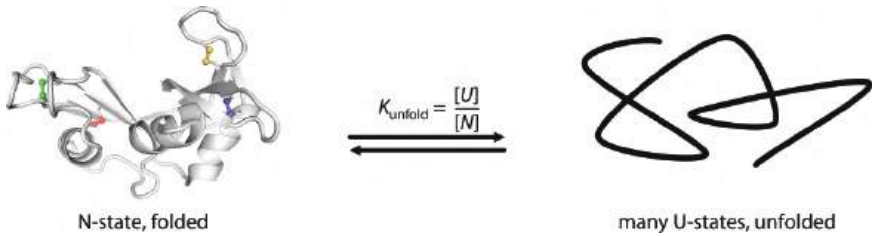


FIGURE 18.56: The two-state model of protein folding. RNase A (PDB-ID 1rat) is shown with its disulfide bonds colored as in Figure 18.54. The ratio of the concentrations of the unfolded states U and the folded state N defines the equilibrium constant K_{unfold} that is dependent on experimental conditions such as temperature or denaturant concentration.

In the two-state model, the thermodynamic equilibrium between native state N and unfolded state U can be described by the equilibrium constant K_{unfold}:

$$K_{unfold} = \frac{[U]}{[N]} = \frac{1-\alpha}{\alpha} \qquad \text{eq. 18.18}$$

Since the total protein concentration is the sum of the concentrations of folded and unfolded protein, K_{unfold} can also be described as a function of the folded fraction α (eq. 18.18). The relation $\Delta G^0 = -RT\ln K_{unfold}$ is used to calculate the free energy of unfolding under standard conditions. A typical two-state unfolding transition, induced by increasing concentrations of denaturants or increasing temperatures, is shown in Figure 18.57.

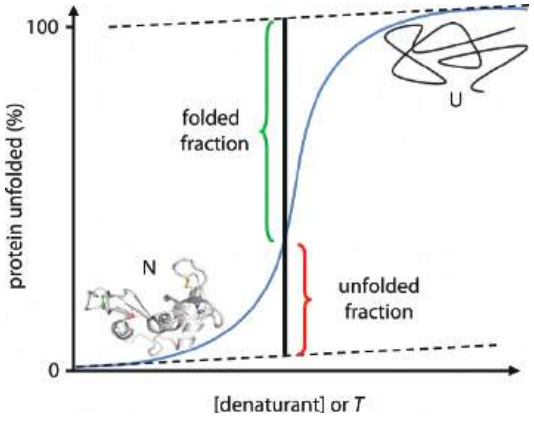


FIGURE 18.57: A two-state unfolding transition. The thermal or chemical denaturation of a protein is followed by a spectroscopic signal. The baselines before and after the transition represent the dependence of the measured signals for the folded and unfolded states on the concentration of denaturant or temperature. By extrapolation of these baselines, the U/N ratios can be determined that correspond to the equilibrium constant K_{unfold} at any given concentration of denaturant or temperature. The inflection point of the transition corresponds to $K = 1$ (50% folded). The slope at the midpoint of the transition is a measure of the cooperativity of unfolding.

The unfolding transition can be monitored by any signal that changes during the experiment, *i.e.* that differs between native and unfolded states of the protein studied, provided that it can be performed under all required experimental conditions (high temperature or denaturant concentrations, extremes of pH). Changes in UV-VIS absorbance (Section 21.2) or fluorescence (Section 21.5) use aromatic amino acids as chromophores and report on changes in tertiary structure. Far-UV circular dichroism (Section 21.3), on the other hand, measures absorption properties of the protein backbone and reports on changes in secondary structure. Nuclear magnetic resonance (Section 22.1) and heat capacity (differential scanning calorimetry, Section 30.2) are also frequently used probes.

Unfolding curves can be divided into three regions, the baseline for the native state at low temperatures or denaturant concentrations, the transition region, and the baseline for the unfolded state at high temperatures or concentrations of denaturant. The baselines represent the change in signal (if any) as a function of temperature or denaturant concentration. They have nothing to do with the actual folding/unfolding event that occurs in the transition region. Each point in the transition region corresponds to the equilibrium between folded and unfolded states. The inflection point is at 50% native/unfolded, which corresponds to $K_{unfold} = 1$ or $\Delta G^0_{unfold} = 0$. The corresponding temperature T_m is the melting temperature of the protein. The middle part of the transition around T_m is almost linear. The steeper this part of the transition, the more cooperative is the (un)folding. A quantitative thermodynamic treatment of protein folding according to eq. 18.18 is only possible if the unfolding is reversible, a prerequisite that is often neglected when protein stability *per se* is the focus (see Box 18.8).

A mathematical description of a reversible two-state unfolding transition was introduced in 1988 by Santoro and Bolen. Denaturant-induced unfolding by urea, guanidinium chloride, or acid typically results in a baseline of the measured signal y. The baselines for the native state N and the unfolded state U are expressed as a linear function of the denaturant concentration $[D]$ with the slope m (Figure 18.57)

$$y_U = y_U^0 + m_U[D] \text{ and } y_N = y_N^0 + m_N[D] \qquad \text{eq. 18.19}$$

Using the law of mass action, we can expand eq. 18.18 to replace the actual concentrations of $[N]$ and $[U]$ with the measured signal y_{obs}.

$$K_{unfold} = \frac{[U]}{[N]} = \frac{y_{obs} - y_N}{y_U - y_{obs}} \qquad \text{eq. 18.20}$$

The free energy of folding ΔG^0_{stab} is the difference between the free energies of the folded (G^0_N) and unfolded states (G^0_U), linearly extrapolated to zero denaturant concentration ($[D] = 0$). The experimentally observed ΔG_{stab} in an unfolding transition is a function of ΔG^0_{stab}, the denaturant concentration, and the cooperativity parameter m_G:

$$\Delta G_{stab} = \Delta G^0_{stab} + m_G \cdot [D] = RT \ln K_{unfold} \qquad \text{eq. 18.21}$$

The linear dependence of unfolded protein on denaturant concentration in the equation above is known as the Tanford model, which applies to many proteins. The cooperativity parameter m_G is defined as

$$m_G = \frac{\partial \Delta G_{stab}}{\partial [D]} \qquad \text{eq. 18.22}$$

The more steeply ΔG_{stab} changes with denaturant concentration, the more cooperative the (un)folding event, and the larger the cooperativity parameter. Solution of eq. 18.20 for y_{obs} and replacement of y_N and y_U (from eq. 18.19) and K_{unfold} (from eq. 18.21) yields the equation needed to describe the observed spectroscopic parameter y_{obs} as a function of denaturant concentration $[D]$ with the fit parameters ΔG_{stab} and m_G (if we fix the parameters for the baselines):

$$y_{obs} = \frac{y_N^0 + m_N[D] - y_U^0 - m_U[D]}{1 + e^{-\frac{\Delta G_{stab} + m_G[D]}{RT}}} + y_U^0 + m_U[D] \qquad \text{eq. 18.23}$$

The procedure to derive the corresponding expression for temperature-induced reversible unfolding events is similar. Here, we start from the Gibbs-Helmholtz equation for protein folding:

$$\Delta G_{stab} = \Delta H_{stab} - T\Delta S_{stab} \qquad \text{eq. 18.24}$$

The enthalpy and entropy terms in eq. 18.24 are temperature-dependent and include the change in heat capacity ΔC_p, which is the difference in heat capacity between the unfolded and the folded states. As derived in Section 1.4 on thermodynamics, we know that

$$\Delta H_{stab}(T) = \Delta H_{stab}(T_m) + \Delta C_p(T - T_m) \qquad \text{eq. 18.25}$$

T_m, the melting temperature of the protein, is taken as the reference temperature. The temperature dependence of the entropy (Section 1.5.3) is

$$\Delta S_{stab}(T) = \Delta S_{stab}(T_m) + \Delta C_p \cdot ln\frac{T}{T_m} \qquad \text{eq. 18.26}$$

At the temperature T_m, we have the same concentrations of native and unfolded protein ($K_{obs} = 1$ and $\Delta G = 0$). We can use this information to eliminate $\Delta S_{stab}(T_m)$ from eq. 18.26.

$$\Delta G_{stab}(T_m) = 0 = \Delta H_{stab}(T_m) - T_m \Delta S_{stab}(T_m) \qquad \text{eq. 18.27}$$

It follows that

$$\Delta S_{stab}(T_m) = \frac{\Delta H_{stab}(T_m)}{T_m} = \frac{\Delta H_m}{T_m} \qquad \text{eq. 18.28}$$

We can now combine eq. 18.24, eq. 18.25, eq. 18.26, and eq. 18.28 to express the free energy ΔG_{obs} as a function the temperature T, the change in heat capacity ΔC_p, and the "melting" enthalpy ΔH_m at the melting temperature T_m.

$$\Delta G_{stab}(T) = -RT \ln K_{unfold} = \Delta H_m \left(1 - \frac{T}{T_m}\right) + \Delta C_p \left(T - T_m - T \cdot ln \frac{T}{T_m}\right) \qquad \text{eq. 18.29}$$

The baselines of the experimentally observed unfolding transition are now a linear function of temperature.

$$y_U = y_U^0 + m_U T \text{ and } y_N = y_N^0 + m_N T \qquad \text{eq. 18.30}$$

Combination of eq. 18.20, eq. 18.29, and eq. 18.30 then yields an expression for the observed signal y_{obs} of the same format as eq. 18.23:

$$y_{obs} = \frac{y_N^0 + m_N T - y_U^0 - m_U T}{1 + e^{-\frac{\Delta H_m}{R}\left(\frac{1}{T} - \frac{1}{T_m}\right) - \frac{\Delta C_p}{R}\left(1 - \frac{T_m}{T} - ln\frac{T}{T_m}\right)}} + y_U^0 + m_U T \qquad \text{eq. 18.31}$$

Eq. 18.23 and eq. 18.31 are the general expressions to describe the signal y_{obs} used to monitor unfolding transitions in order to extract the thermodynamic parameters of (un)folding. In practice, temperature-induced unfolding experiments have the advantage that the measured data are directly related to thermodynamic quantities (ΔH, ΔG, ΔC_p). Chemically induced unfolding introduces additional parameters such as the dependence of ΔG on the denaturant concentration, but the thermodynamic stability is extrapolated to the absence of denaturant.

The plot of ΔG_{stab} as a function of temperature, termed the protein stability curve, has a bell shape (see Figure 1.26). The overall curvature of the stability curve, and therefore the temperature range where the protein is stable, is determined by the values of ΔC_p and ΔH_m. The origin of ΔC_p is mostly related to the hydrophobic effect and comes from the increase in solvent-exposed non-polar surface area when the protein unfolds. ΔC_p is usually positive in the range 0–30 kJ mol^{-1} K^{-1}, and lies at around 8 kJ mol^{-1} K^{-1} for most proteins. The melting enthalpy ΔH_m is also positive and in the range between almost zero to a few hundred kJ mol^{-1}. From Figure 18.58, we see that ΔG_{stab} reaches zero at two different temperatures. This shape predicts that proteins actually have two melting points, one at low temperature, one at high temperature. We have treated these processes, cold denaturation and heat denaturation, in Section 1.6.2. Cold denaturation is enthalpy-driven, while heat denaturation is entropy-driven (see Box 1.9).

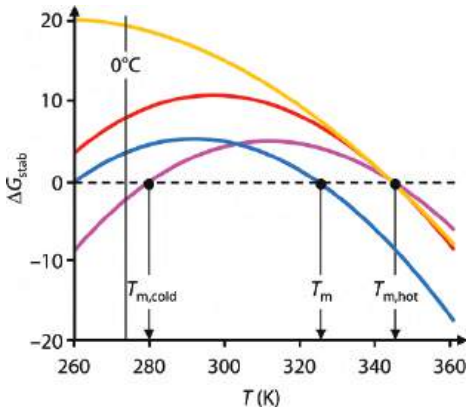


FIGURE 18.58: Protein stability curves. The freezing point of water is marked by the line at 0°C (273 K). A mesophilic (blue curve) protein is shown with a T_m of 323 K. The T_m for its cold denaturation is at 260 K, below the freezing point of water. The three stability curves of thermostable proteins (magenta, red, and orange) have the same $T_{m,\,hot}$ for heat denaturation at 343 K but different $T_{m,\,cold}$ for cold denaturation. A protein can be stabilized in several ways: the magenta curve has the same ΔH_m and ΔC_p as the mesophilic protein but a 20°C higher T_m and is just right-shifted from the mesophilic curve. The shift makes the cold denaturation temperature $T_{m,\,cold}$ at 280 K experimentally accessible. The red curve has a more positive ΔH_m but the same ΔC_p as the magenta and blue curves. The orange curve has a smaller ΔC_p than the other curves.

Cold denaturation was predicted from eq. 18.29, but was not observed experimentally for quite some time. This is because for mesophilic proteins, the T_m for cold denaturation is typically below the freezing point of the buffer system, making it difficult to monitor with most biophysical techniques. Stability curves for thermophilic proteins are often shifted to higher temperature, which can shift T_m for cold denaturation into the experimentally accessible range (Figure 18.58).

The extraction of thermodynamic parameters from protein unfolding experiments according to the two-state model is only possible if the unfolding is reversible. If this is not the case, the midpoint of the transition still provides information about the stability of the protein against the denaturant. In comparative studies of protein stabilities, it is often enough to measure the unfolding transition and compare the values of their midpoints. Such comparative studies are often performed using a *Thermal Shift Assay* (TSA; Box 18.8).

BOX 18.8: THERMAL SHIFT ASSAY FOR PROTEIN STABILIZATION AND LIGAND BINDING.

The exposure of hydrophobic surfaces upon temperature-induced protein unfolding can be used to estimate the relative stabilities of protein variants or to rank the affinities of ligands that bind to, and hence stabilize, the protein (Box 18.9). Unfolding leads to exposure of hydrophobic surfaces. The hydrophobic dye SYPRO Orange binds to these surfaces, which leads to an increase in its fluorescence (Section 21.5). The fluorescence increase thus reports on the progress of unfolding. The measurement can be automated and conducted in micro-titer plates such that dozens of protein variants or protein-ligand complexes can be ranked in a matter of a few hours. Protein-ligand complexes are always more stable than the apo-protein. Single amino acid exchanges, on the other hand, may have a stabilizing or destabilizing effect. Higher stability will lead to a higher melting temperature T_m, and *vice versa* will a destabilizing mutation lead to a lower T_m. The magnitude of the shift in T_m is a measure of the extent of protein stabilization by ligands or stabilization/destabilization by mutations (Figure 18.59).

(*Continued*)

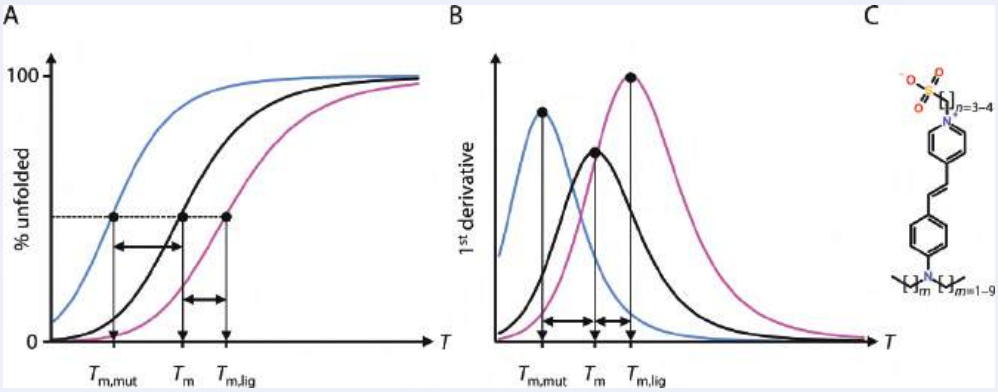


FIGURE 18.59: Thermal Shift Assay. A: The melting temperature is the inflection point of the transitions. T_m for the reference (wild-type or apo-form) is colored black. A positive shift in T_m to $T_{m,lig}$ is the readout for stabilization of the protein by binding of a ligand. A negative shift to $T_{m,mut}$ is possible for mutations that destabilize a protein. B: Alternatively, the experimental data can be plotted as the first derivative. T_m is then identified as a minimum or maximum, depending on whether the signal decreases or increases, and the method is called Differential Scanning Fluorimetry (DSF). C: The general structure of SYPRO Orange dyes. In TSAs, the dye with $n = 3$, $m = 5$ is often used.

We have discussed at the beginning of this Chapter that a protein can be stabilized by stabilizing the native state or by destabilizing the unfolded state. It may not be immediately obvious that the stabilization of the folded state of the protein is an inevitable consequence of ligand binding. In fact, any ligand that binds more strongly to the native than to the unfolded state of a protein causes an increase in protein stability (Box 18.9).

BOX 18.9: LIGAND BINDING INEVITABLY STABILIZES THE NATIVE STATE OF A PROTEIN.

We consider a ligand L that can in principle bind both to the native state N and to the unfolded state U of a protein. The thermodynamic cycle describing this system is given by

$$\begin{array}{ccc} N & \underset{K_{fold}}{\rightleftharpoons} & U \\ K_{d,N} \updownarrow & & K_{d,U} \updownarrow \\ NL & \rightleftharpoons & UL \end{array} \qquad \text{scheme 18.2}$$

The equilibrium constant K_{fold} for protein folding is

$$K_{fold} = \frac{[N]}{[U]} \qquad \text{eq. 18.32}$$

and the equilibrium dissociation constants of the ligand complexes are

$$K_{d,N} = \frac{[N][L]}{[NL]} \text{ and } K_{d,U} = \frac{[U][L]}{[UL]} \qquad \text{eq. 18.33}$$

We define an apparent equilibrium constant K_{app} (see Section 2.4) as the ratio of all folded and all unfolded species, irrespective of whether the ligand is bound:

(*Continued*)

$$K_{app} = \frac{[N]+[NL]}{[U]+[UL]} \qquad \text{eq. 18.34}$$

Substituting $K_{d,U}$ and $K_{d,N}$ into the definition for K_{app} gives

$$K_{app} = \frac{[N]+\frac{[N][L]}{K_{d,N}}}{[U]+\frac{[U][L]}{K_{d,U}}} = \frac{[N]}{[U]} \cdot \left(\frac{1+\frac{[L]}{K_{d,N}}}{1+\frac{[L]}{K_{d,U}}} \right) = K_{fold} \cdot \left(\frac{1+\frac{[L]}{K_{d,N}}}{1+\frac{[L]}{K_{d,U}}} \right) \qquad \text{eq. 18.35}$$

If the ligand stabilizes the native state, then the condition

$$K_{app} > K_{fold} \qquad \text{eq. 18.36}$$

must be true. This means that

$$1+\frac{[L]}{K_{d,N}} > 1+\frac{[L]}{K_{d,U}} \qquad \text{eq. 18.37}$$

which it true for

$$K_{d,U} > K_{d,N} \qquad \text{eq. 18.38}$$

If $K_{d,N}$ is smaller than $K_{d,U}$, then the ligand binds stronger to the native state than to the unfolded state. Eq. 18.38 thus demonstrates that binding of any ligand that interacts more strongly with the native state N than with the unfolded state U, stabilizes the native state of the protein. This is true for all ligand concentrations. If a "ligand" binds more strongly to the U-state, the protein is destabilized. In this case the ligand is called a denaturant.

18.5.5 Folding pathways and mechanisms of protein folding

So far, we have looked at protein folding mostly from the thermodynamic point of view. For the complete description of folding pathways, we also need to know all states populated, and the microscopic rate constants for their inter-conversion. Protein folding pathways have been elucidated for a number of proteins, including ubiquitin, lysozyme, RNase, myoglobin, and the cold shock protein CspB. The best understood examples in terms of kinetics and thermodynamics are small, autonomously folding single-domain proteins that display a two-state folding transition without folding intermediates (Figure 18.57). However, this simple pathway seems to be the exception rather than the rule. Many proteins populate one or more folding intermediates, use different folding pathways, or fall into kinetic traps from which they need to be rescued. Folding of proteins always competes with protein aggregation (Figure 18.60). Furthermore, protein folding *in vivo* occurs in a completely different context to protein folding in *in vitro* experiments: while *in vitro* experiments typically start from a complete polypeptide chain in the unfolded state, folding *in vivo* starts as soon as a nascent polypeptide exits the ribosome, and may follow an entirely different folding pathway.

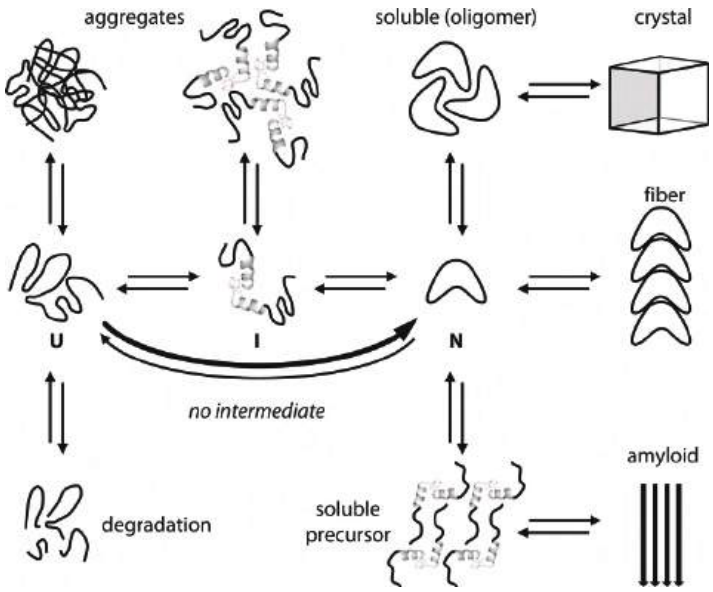


FIGURE 18.60: Overview of protein folding, assembly, and aggregation. The result of ribosomal peptide synthesis is an unfolded protein U, which can either fold directly or *via* intermediates (labeled I) to the native state N. The native state can proceed to assemble into oligomers, fibers, and sometimes even crystals. Folding-incompetent proteins, such as variants or protein fragments from premature translation termination, are degraded. Chaperones are vital for the folding of large multi-domain proteins and to avoid aggregation. ΔG_{fold} is small for most proteins. The smaller ΔG_{fold}, the more populated the unfolded state in equilibrium. The unfolded state can then serve as a starting point for the (mis-)folding into soluble precursors of toxic amyloids (Section 18.5.6).

Several models have been proposed to explain how proteins can fold in a matter of milli- or sometimes microseconds, and most of these models separate secondary structure formation from tertiary structure formation. The *nucleation model* assumes that a group of residues close in sequence forms a secondary structure element (helix, loop, or β-hairpin), from which new secondary structure is propagated along the chain. Tertiary structure forms from interactions of the string of secondary structure elements along the polypeptide chain. By contrast, in the *diffusion-collision model*, secondary structure elements form independently at several locations throughout the polypeptide. These elements then diffuse, collide, and bind to each other until the polypeptide chain coalesces into the native tertiary structure. Finally, the *hydrophobic collapse model* proposes that the hydrophobic effect drives compaction of the polypeptide chain, which then enables formation of secondary structure. A hydrophobic collapse imposes steric restrictions on movements, and also much reduces the number of water molecules around the polypeptide chain that compete with NH- and CO-groups for hydrogen bonds. Limited movement and desolvation thus promote secondary structure formation.

While the nucleation and diffusion-collision models propose that secondary structure forms before protein compaction by tertiary structure formation, the hydrophobic collapse model states the reverse: the three-dimensional collapse precedes secondary structure formation. The *nucleation-condensation model* is a synthesis of both models. According to this model, secondary and tertiary structure is formed in parallel: a nucleus of neighboring residues forms a native-like secondary structure that is subsequently stabilized by tertiary interactions.

In reality, different proteins will follow different folding pathways that bear more or less similarity to one or the other model. For example, helical membrane proteins are thought to fold sequentially, following the nucleation and the diffusion-collision models. First, α-helices are formed within the membrane, which is a thermodynamically favored event in the absence of water (Section 18.4.4). The helices then diffuse laterally to form the helical bundles of the native structure. By contrast, small globular proteins such as SH3 domains and the chymotrypsin inhibitor fold according to the nucleation-condensation model, with parallel formation of secondary and tertiary structure.

The different steps of protein folding occur on very different time scales (Table 18.12). The fastest steps in folding are formation of secondary structure elements and the hydrophobic collapse of the polypeptide chain to generate a flexible but compact, *hydrated intermediate* known as a *molten globule*. While for proteins following a two-state model the molten globule represents a transition state that cannot be observed, it is an experimentally tractable folding intermediate for oligomeric and multi-domain proteins. Molten globules have substantial native-like secondary structure content with a topology close to the native state, but they still lack many specific tertiary interactions and are thus not functional (enzymes are inactive). The key stabilizing interactions of molten globules are non-specific hydrophobic interactions between non-polar side chains that are concentrated at the core of the molten globule. Water is present throughout the entire structure, which results in a Stokes radius (see Section 29.2.2) that is 10–15% larger than that of the corresponding folded protein. During later stages of folding the water is completely excluded from the hydrophobic core, and the tertiary interactions characteristic for the folded state are established. In multi-domain proteins, each domain may fold individually directly or *via* a molten globule and then associate with other domains to form the native state. For oligomeric proteins, the folding of individual domains or subunits is followed by subunit assembly. Subunit assembly proceeds *via* *n*-mer intermediates, not sequentially by addition of one monomer to another (Scheme 18.3). A tetramer T, for instance, will have dimers (D) along its folding pathway, possibly with different conformations (D') than those present in the folded tetramer.

$$4\,M_u \rightleftharpoons 4\,M_f \rightleftharpoons 2\,D' \rightleftharpoons 2\,D \rightleftharpoons T' \rightleftharpoons T$$ scheme 18.3

Similarly, hexamers, dodecamers, and 24-mers have dimers, trimers, tetramers, and hexamers as intermediates. Folding and assembly of oligomers bears the danger of aggregation, which at higher monomer concentrations can efficiently compete with oligomerization.

The sequence secondary-tertiary-quaternary structure formation during folding sounds logical, but is by no ways strict. Often these stages are inter-dependent: the hydrophobic collapse of secondary structure elements drives further secondary structure formation by exclusion of water, and reshuffling of secondary structure elements is possible even after the quaternary structure has been established. For multi-domain proteins, monomer folding and assembly can occur simultaneously.

TABLE 18.12:
Time scales for molecular processes in protein folding.

time range	extent	scale	action
fs – ps	< 10 pm	local	bond vibrations (stretch and bend); source of IR signal; nucleobase wobbling
~ 1 ps		local	rotation correlation time of water molecules; flipping of aromatic rings; sugar puckering
< 1 ns		local	free side-chain rotation and lifetime of hydrogen bonds
1 ns – 1 µs	< 0.5 nm	medium	movement of protein loops; quasi-harmonic twisting of β-sheets; flexing of α-helices
many ns – 1 µs	a few nm	medium	small peptide folding and helix/coil transition
many µs – 1 ms	< 1 nm	medium	subunit (allostery) and domain (hinge) motions
1 µs – 1 ms	> 1 nm	large	membrane transfer of proteins
ms – many min	> 10 nm	large	protein subunit association
1 ms – > 1 s	> 10 nm	large	protein folding and unfolding

In the following, we will take a closer look at the fast and slow steps (Section 18.5.5.1 and 18.5.5.2), and at kinetics of protein folding (Section 18.5.5.3).

18.5.5.1 Fast steps in protein folding: secondary structure formation

The formation of α-helices in solution has been studied in considerable detail and reveals many of the key requirements for any secondary structure to form. An isolated α-helix in aqueous solution is in rapid exchange with the unfolded (random coil) state, the conversion occurs with rate constants of about 10^5–10^7 s^{-1}. While the rate constant for helix formation is not dependent on the final length of the helix, the unfolding of helices is strongly length-dependent: the longer the helix, the slower the unfolding.

Helix formation may be separated into an initiation and an extension phase. Thermodynamically, the entropic penalty for helix initiation is much higher than for helix extension. For helix initiation, *i.e.* formation of the first turn, the dipoles of four peptide bonds must align parallel to each other, which is an improbable event. Furthermore, the dipoles are placed side-by-side, without a displacement along the yet to be established helix axis, which is not an energetically favorable arrangement (see Figure 17.8). By contrast, helix extension after the first turn requires alignment of just one additional peptide bond per residue and the dipoles now interact favorably in a head-to-tail orientation, stabilizing the folding helix and creating the helix macro-dipole.

Two models have been developed to explain the kinetics of helix-random coil interconversion. The *Zimm-Bragg model* (see Section 3.3.1) looks at formation of the first hydrogen bond in the helix followed by helix extension, whereas the *Lifson-Roig model* defines helicity by ϕ/ψ angles, not by hydrogen-bonding. In both models, the rate-limiting step is helix initiation, which can in principle occur anywhere in the sequence but is slower than helix extension by a factor of 300–500. The slow initiation points to a high activation barrier, which can be explained by the fact that at least five residues must adopt α-helical conformation to form just a single *i*,*i*+4 hydrogen bond that spatially fixes the intervening three residues. By contrast, helix extension only requires the fixing of the conformation of one additional residue at a time, and this residue is the next in the sequence and thus close by.

In the Zimm-Bragg and Lifson-Roig models, all amino acids ideally have the same probability of incorporation into the helix. However, the propensities of individual residues to form α-helices are different (see Table 18.5). For example, the charged amino acids Glu and Asp interact with the helix dipole, which makes them about four times more likely to extend the helix at the N-terminus than at the C-terminus. Polar side chains such as Ser, Thr, and Asn can hydrogen-bond to the backbone and thus interfere with helix elongation. Also, the growing α-helix zips up in both directions until a breakpoint, such as Pro or a capping residue (Gly as a C-cap for example), is reached. At this point, the probability of incorporation increases for any residue at the opposing end of the growing α-helix.

The zipper-like extension of secondary structure also applies to super-secondary structures such as the helical coiled-coils and the collagen triple helices. β-sheets also form in a similar manner to helices. They are initiated by a minimum of three residues in the extended conformation (see Figure 18.27) and then extend rapidly in both directions until a break is reached, *e.g.* Pro. It has been suggested that turns (see Figure 18.33) are important decision points in protein folding because they can define the direction that the polypeptide chain takes. The strongly preferred residues Pro and Gly in turns appear to slow β-sheet extension and thus have strong guidance for folding of the entire protein.

18.5.5.2 Rate-limiting steps and protein folding *in vivo*

Among the slowest steps in protein folding are prolyl isomerization, disulfide bond formation, and folding of large domains that are prone to aggregate. *In vivo*, these reactions are catalyzed by specific sets of enzymes, the *prolyl isomerases* (PPIs), *protein disulfide isomerases* (PDIs), and *chaperones*. We will only outline the principles of action of these enzymes.

Although ribosomal protein synthesis defaults to the *trans*-configuration, *cis*-peptide bonds (see Figure 18.4) are quite common in proteins (5–6%). Many proteins containing immunoglobulin domains, including antibodies and cell surface receptors, have *cis*-prolyl peptide bonds. The equilibrium is established slowly at physiological pH: *cis*-peptides form with rate constants of only 0.1–0.01 s^{-1}. Spontaneous prolyl isomerization can therefore be the rate-limiting step during folding of proteins that contain *cis*-prolyl peptide bonds in the folded state (see Box 28.1). Slow prolyl isomerization can have biological relevance and can act as switch and timer (Box 18.10).

PPIs are very efficient enzymes with k_{cat}/K_M values of 10^5–10^6 $M^{-1}s^{-1}$, accelerating the *trans*/*cis* isomerization equilibrium by a factor of $> 10^6$. Proteins with *cis*-prolyl peptide bonds fold faster in the presence of PPI, because PPIs accelerate this rate-limiting step. Folded forms of proteins with non-native *trans*-prolyl peptide bonds do exist and have a native-like structure, and in the case of enzymes these may have partial activity. An example is RNase T1, which has four Pro residues and two *cis*-prolyl peptide bonds in its native state. A folding intermediate with a *trans*-prolyl peptide bond at Pro93 has a compact native-like structure as detected by NMR hydrogen exchange (Section 18.5.5.3) and it has some RNase activity. The complete folding scheme of RNase T1 encompasses all 16 possible *cis/trans* isomers.

BOX 18.10: *CIS/TRANS* PROLYL ISOMERIZATION AS A MOLECULAR SWITCH AND TIMER.

Spontaneous, non-catalyzed *cis*/*trans* peptide bond isomerization plays a role in viral infection where it acts as a molecular timer and switch. The *E. coli* filamentous phage fd carries the gene-3 protein (G3P) on its tip. G3P consists of two interacting domains that bury between them an unusual Gln212-Pro213 *cis*-peptide bond. The bond is unusual because normally the preceding residues in *cis*-prolyl bonds have aromatic side chains (His, Phe, Tyr, Trp) that pack on top of the proline to at least somewhat stabilize the unfavorable *cis*-configuration (Section 18.1.2). The Gln-Pro *cis*-peptide in G3P has high energy but cannot relax to the *trans*-configuration because it is locked by the inter-domain interaction. With the lock in place, the phage is in a robust but non-infectious state. Once G3P binds to the F-pilus of *E. coli*, the domains in G3P come apart and the phage can now become activated for infection by the spontaneous conversion of the *cis*-peptide to *trans*. The isomerization sets both a molecular timer and flips a structural switch in G3P: the accompanying large conformational change in the backbone exposes a new binding site on G3P for the *E. coli* TolA protein, the ultimate fd phage receptor. The sequence of events during infection takes a few minutes and needs to be completed before the Gln-Pro *trans*-peptide spontaneously converts back into *cis*. This back-conversion is slow enough for the infection to finish but quickly enough to revert the virus back into an inactive state should infection be unsuccessful. The time constant, or lifetime of the "on"-state is encoded in the protein sequence around Pro213: mutation of Gln212 to other residues decreases the time constant, and the virus reverts to an inactive state before infection is complete.

The conformational switch enabled by prolyl isomerization has also been implicated in the regulation of intracellular processes such as transmembrane signaling, protein phosphorylation and the cell cycle. In contrast to extracellular regulation, where prolyl isomerization is un-catalyzed, intracellular processes can in addition be regulated by PPI. An example is Pin1, a phosphorylation-dependent isomerase that acts on phospho-Ser/Thr-prolyl peptide bonds and thus couples prolyl isomerization to the tightly regulated activities of kinases and phosphatases.

We have encountered the important role of disulfide bond exchange for protein folding when discussing the Anfinsen experiment (Section 18.5.2). The number of possible disulfide pairings increases with the number of Cys residues. If not all Cys are tied up in disulfide bonds, the number of possible combinations becomes even higher. Only a single set of disulfide bonds corresponds to the native state. PDI both facilitate the (*de novo*) introduction of disulfide bonds and the exchange of incorrectly formed disulfide bonds. In eukaryotes disulfide bond formation takes place in the endoplasmic reticulum, which has a less negative redox potential (–190 mV) than the cytoplasm (–230 mV), and thus facilitates Cys oxidation (see Section 5.3). PDIs display broad substrate specificity, and have no sequence specificity. A low specificity ensures that they can assist a broad spectrum of

proteins in folding. The active sites of PDIs contain a reactive pair of Cys in a CXXC sequence that forms the first turn of a long α-helix. The influence of the helix dipole causes unusually low pK_A values of these Cys residues. Introduction of disulfide bonds starts from the oxidized PDI, leaving the PDI in a reduced state. In eukaryotes, the reduced PDI is re-oxidized by endoplasmic reticulum oxidoreductin 1 (Ero1). Proteins that spontaneously fold into a near-native state will bring the correct Cys residues together, so only oxidation needs to be catalyzed by PDI. For incorrectly formed disulfide bonds, catalysis of disulfide exchange can proceed *via* all possible intermediates, regardless of whether they involve conformational changes in the protein substrate or not (Figure 18.61).

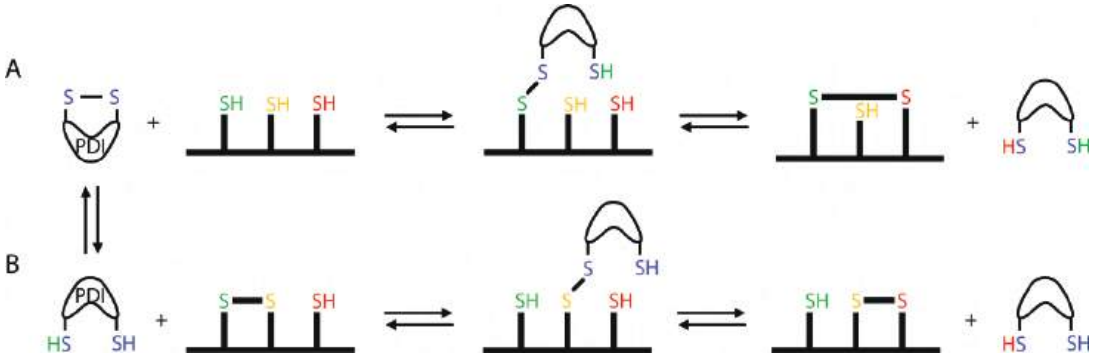


FIGURE 18.61: Disulfide bond formation and exchange during protein folding. A: Disulfide bond formation at the expense of oxidized PDI. Oxidized PDI forms a mixed disulfide bond in an enzyme-substrate intermediate. Intramolecular disulfide bond formation during protein folding releases reduced PDI. B: PDI-catalyzed disulfide exchange *via* a mixed disulfide intermediate. PDI-catalyzed disulfide bond formation and exchange (shuffling) can accelerate the folding by a factor of 100. In bacteria, separate enzymes are used for introduction and shuffling of disulfide bonds.

"Chaperones" is the collective term for a family of unrelated proteins that aid the folding and assembly of other proteins in an ATP-dependent fashion. Chaperones act as "catalysts" by accelerating the folding of their substrates but they also shift the equilibrium towards the folded state by expending energy in the form of ATP hydrolysis. Proteins that are associated with chaperones and regulate their activity but that do not hydrolyze ATP are termed co-chaperones. Chaperones bind to the solvent-exposed hydrophobic surfaces typical for unfolded proteins and folding intermediates in order to suppress aggregate formation (Section 18.5.6).

The main chaperone systems, Hsp60 and Hsp70, have been defined based on the molecular mass of their heaviest component. Heat shock proteins (Hsps) are chaperones produced as a stress response at elevated temperatures. GroEL, a tetradecamer of two seven-membered rings stacked on top of each other (Figure 18.38), is the main chaperone of the Hsp60 group in *E. coli*. Together with its co-chaperone GroES, GroEL cordons off a large hydrophobic space to isolate aggregation-prone unfolded proteins from the remainder of the cellular proteins. In an ATP-dependent reaction, the GroEL-GroES complex releases the folding intermediate inside the complex for another attempt at folding, *i.e.* to lift the folding intermediate out of its local energetic minimum. If the protein has folded, it will have polar surfaces with low affinity for the GroEL-GroES complex, and be released. Iterative rounds of assisted folding may be required until the native state is reached. If not, the protein is eventually degraded by proteases, many of which (Lon, Clp, HtrA) are also heat shock proteins.

The *E. coli* chaperone DnaK is a representative of the Hsp70 family of chaperones that acts in conjunction with its co-chaperones DnaJ and GrpE. Similar to GroEL, DnaK binds and hydrolyzes ATP. DnaJ stimulates the ATPase activity, and GrpE is a nucleotide exchange factor that accelerates ADP/ATP exchange after hydrolysis. The ATP state of DnaK binds hydrophobic regions rapidly, but with low affinity. Hydrolysis of ATP to ADP and phosphate generates a high-affinity state, and the peptides remains stably bound until the ADP is exchanged for ATP. Hsp70 proteins promote assembly and disassembly of oligomers, and keep newly synthesized proteins in an unfolded state that can be translocated through membranes.

18.5.5.3 Kinetics of protein folding: the two-state model revisited

In Section 18.5.4, the two-state model for protein folding, with the native state N and the unfolded state U, was treated from the equilibrium point of view. Let us now consider the same system from a kinetic perspective. The equilibrium between the N and U states is characterized by the microscopic rate constants k_{fold} and k_{unfold} (Scheme 18.4).

$$N \underset{k_{fold}}{\overset{k_{unfold}}{\rightleftharpoons}} U$$

scheme 18.4

The description of a reversible reaction according to Scheme 18.4 is discussed in detail in Section 8.1. Here, we apply this knowledge to protein folding. The equilibrium constant K_{unfold} is described by the ratio of the folded and unfolded protein concentrations or by the ratio of the rate constants:

$$K_{unfold} = \frac{[U]}{[N]} = \frac{k_{unfold}}{k_{fold}} \qquad \text{eq. 18.39}$$

When we start from a solution of a native protein and rapidly dilute it into a high concentration of denaturant, we can follow the unfolding reaction as a function of time, using any of the spectroscopic probes that are also suitable for measuring equilibrium protein unfolding. Similarly, we can follow the folding reaction if we start from a solution of unfolded protein that we rapidly dilute into native conditions. In either case, the observed rate constant k_{obs} at each denaturant concentration will be an apparent rate constant that is the sum of k_{fold} and k_{unfold} at that particular denaturant concentration:

$$k_{obs} = k_{fold} + k_{unfold} \qquad \text{eq. 18.40}$$

k_{obs} is obtained by fitting an exponential function (Chapter 7) to each unfolding trace. Plotting the logarithms of the k_{obs} as a function of denaturant concentration results in a V-shaped plot that has been named the *Chevron plot* (Figure 18.62), from which the microscopic rate constants k^0_{fold} and k^0_{unfold} in the absence of any denaturant can be extracted.

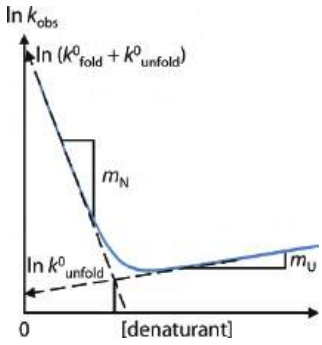


FIGURE 18.62: Chevron plot. A plot of the logarithm of the apparent rate constant of folding ln(k_{obs}) as a function of denaturant concentration has a V-shape. Extrapolations of the linear regions to zero denaturant yield the logarithms of the microscopic rate constants k^0_{fold} and k^0_{unfold}. The intersection of the linear extrapolations is the denaturant concentration at the midpoint of the transition.

The Chevron plot is V-shaped because both microscopic folding and unfolding rate constants are affected by the denaturant, an effect that has nothing to do with the folding process itself. Specifically, the logarithms of the rate constants k_{fold} and k_{unfold} change linearly with the denaturant concentration:

$$\ln k_{fold} = \ln k^0_{fold} - m_N \cdot [D] \qquad \text{eq. 18.41}$$

$$\ln k_{\text{unfold}} = \ln k^0_{\text{unfold}} + m_{\text{U}} \cdot [D] \qquad \text{eq. 18.42}$$

The values of the slopes m_{N} and m_{U} describe how strongly the rate constants depend on the denaturant D. The intersections of these lines with the *y*-axis are the logarithms of the microscopic rate constants in the absence of denaturant. If we combine equations eq. 18.40–eq. 18.42, we obtain the general expression for the Chevron plot.

$$\ln k_{\text{obs}} = \ln\left(k_{\text{fold}} + k_{\text{unfold}}\right) = \ln\left(k^0_{\text{fold}} \cdot e^{-m_{\text{N}} \cdot [D]} + k^0_{\text{unfold}} \cdot e^{m_{\text{U}} \cdot [D]}\right) \qquad \text{eq. 18.43}$$

For low denaturant concentrations, $\ln(k_{\text{obs}})$ approaches $\ln(k^0_{\text{fold}} + k^0_{\text{unfold}})$, meaning the sum of the microscopic rate constants is obtained from the *y*-axis intercept. For high denaturant concentrations, $\ln k_{\text{obs}}$ is dominated by $\ln(k^0_{\text{unfold}} \cdot e^{m_{\text{U}} \cdot [D]})$, which is equal to $\ln(k^0_{\text{unfold}}) + m_{\text{U}} \cdot [D]$. The microscopic rate constant for folding may be obtained from extrapolation of the linear dependence of k_{obs} on [*D*] to the *y*-axis (Figure 18.62).

Rate constants for protein folding and unfolding are typically determined in experiments that jump from native to unfolding conditions or *vice versa*. The only requirement for these jump-techniques to work is that the folding/unfolding is reversible under all conditions. To initiate folding reactions, solutions of the unfolded proteins can be diluted into buffer with low denaturant concentrations. Inversely, unfolding can be initiated by diluting the native proteins into buffer with a high denaturant concentration. Slow folding reactions with half-lives of > 10 s can be initiated by manual mixing and then be followed in standard spectrometers. Fast reactions, however, require stopped-flow (Section 28.1) or quench flow systems (Section 28.2) to rapidly initiate and follow the reactions on the millisecond time scale.

With relaxation methods, reactions with half-lives in the microsecond range can be monitored (see Box 28.3). These techniques cause a shift in the equilibrium constant by a rapid change in temperature (T-jump) or pressure (p-jump). The equilibrium constant for protein folding is pressure-dependent; the unfolded state has a larger volume than the native state. According to the principle of Le Chatelier, an increase in pressure induces folding, while a reduction promotes unfolding. In relaxation methods, the relaxation of the system to the new equilibrium with $k = k_{\text{fold}} + k_{\text{unfold}}$ is monitored.

Using transient kinetic methods (Chapter 28), the rapid generation of secondary structures and the slower subsequent steps of prolyl isomerization, disulfide bond formation, etc. can be measured. If the rate constants of each step are sufficiently different, the folding reaction will have several phases that can be evaluated individually by a multi-exponential curve fit. Multi-exponential folding is common for proteins that require the slow *cis/trans* isomerization of peptide bonds. Due to the low difference in free energy between *cis*-prolyl and *trans*-prolyl peptide bonds, both configurations are populated in the unfolded state. If the native state contains a *cis*-prolyl peptide bond, the folding kinetics will show a fast phase (*burst phase*) that corresponds to folding of those polypeptides that already have the correct *cis*-peptide, and a slow phase corresponding to those that need to first isomerize to the *cis*-peptide before fast folding can take place.

While transient kinetic methods using optical spectroscopy as a probe monitor global folding of the protein, *local protein folding* can be monitored by NMR spectroscopy in H/D exchange experiments (see Section 22.1.7). To do so, the unfolded protein is diluted from a denaturing buffer based on H_2O into a native buffer containing D_2O. Amide protons (H-atoms) from the peptide groups can be replaced by deuterons (D-atoms) from the new solvent. Deuterons are NMR-silent. Therefore, the exchange can be followed as a decrease in intensity of resonances from amide protons as a function of time. The velocity of the exchange depends on the local accessibility of the protein backbone. The amide protons of solvent-exposed peptide groups exchange rapidly, while amide protons of buried peptide groups do not show exchange or exchange only very slowly with deuterons (see Section 22.1.7). Thereby, it is possible to determine when in the folding process each amino acid becomes protected from the solvent. This technique is particularly suited to delineate folding pathways of proteins folding through one or more intermediates.

18.5.6 Protein folding diseases

Proteins with a small negative value for ΔG_{fold} are prone to unfold at slightly elevated temperatures. Mutations in proteins can make ΔG_{fold} less negative, so that a substantial fraction of the protein may be in the unfolded state at physiological temperatures. The presence of unfolded proteins in the cell is often linked to disease. There are two possible reasons for these detrimental effects: loss of function of the folded state or gain of function of the unfolded state. The diseases are more severe in homozygous patients where both copies of the affected genes are mutated, and only the destabilized protein variant is synthesized.

Loss of function of the folded state often results from a mutation in the hydrophobic core that destabilizes the whole protein. An example is the Leu444Pro mutation in glucocerebrosidase, which leads to Gaucher disease. When glucocerebrosidase is mutated (or absent), its substrates (glycosphingolipids) accumulate in the lysosome. Together with Fabry disease and Tay-Sachs disease, where degradation of glycosphingolipids is also impaired, Gaucher disease belongs to the class of *lysosomal storage diseases*. There are about 200 mutations known to cause Gaucher disease, most of which affect folding of glucocerebrosidase to different extents, which in turn leads to differently severe outcomes.

The tumor suppressor p53 is a transcription factor that can activate DNA repair, initiate apoptosis, and arrest the cell cycle. p53 is involved in so many cellular processes that it has been dubbed the *guardian of the genome*. Its marginal thermodynamic stability, with a melting temperature of around 45°C, is beneficial for rapid regulation of p53 activity. The downside is that mutations readily lead to temperature-sensitive variants of p53 that increase the population of the unfolded state at the physiological temperature of 37°C. About 30% of all cancer-related p53 mutations generate temperature-sensitive variants. In principle, it is possible to stabilize these slightly destabilized mutated proteins with a ligand that binds to the folded state of p53 (Box 18.9), which would be an approach toward treatment of these diseases. This has been attempted for the cancer-causing Y220C mutation in p53 where a cavity on the surface of p53 created by the mutation was filled by a small molecule.

A third example of folding diseases is caused by mutations that disrupt the collagen triple helix (Figure 18.26). *Osteogenesis imperfecta* is a disease caused by mutations in collagen that hinder the close approach of the three strands necessary for forming the triple helix. In principle, scurvy is also a type of a collagen folding disease. A lack of vitamin C leads to insufficient Pro-hydroxylation and destabilization of the triple helix by fewer inter-strand crosslinks.

Gain of function of the unfolded state is observed when unfolded proteins or their degradation products adopt new functions. If the concentration of an unfolded protein or one of its degradation products reaches a critical level, it may aggregate into regular structures and form *amyloid plaques*. In Alzheimer's disease, for instance, the combined action of the two proteases BACE-1 and γ-secretase on the amyloid precursor protein APP liberates a 4 kDa peptide of just forty residues termed Aβ. The peptide can assemble into amyloids, which are infinite β-sheets with a superhelical twist. The plaques have been implicated in Alzheimer disease progression. A few more examples are collected in Table 18.13.

TABLE 18.13:
Examples of protein folding diseases

disease, protein	remarks
hypercholesterolemia, low-density lipoprotein receptor LDLR	LDL is a major cholesterol-carrying lipoprotein. Non-functional LDLR (> 200 mutations known) leads to high blood levels of cholesterol.
cystic fibrosis. cystic fibrosis transmembrane regulator	Integral membrane Cl^- channel. ~1500 gene mutations known, ⅔ of all patients have ΔF508.
phenylketonuria. phenylalanine hydroxylase	> 400 gene mutations known.
Marfan syndrome, fibrillin	Fibrillins are part of the connective tissue between cells and regulate transforming growth factor beta (TGF-β) bioavailability. Mutants do not bind to TGF-β.
osteogenesis imperfecta and scurvy, procollagen	see text
sickle cell anemia, hemoglobin	$\alpha_2\beta_2$ heterotetramer. The β-chain Glu6Val mutant polymerizes at low pO_2.
Tay-Sachs disease, β-hexosaminidase	Lysosomal storage disease. The enzyme can no longer metabolize GM2 gangliosides.
retinitis pigmentosa, rhodopsin	> 100 gene mutations known to cause this autosomal dominant disease.
cancer, p53	see text
Alzheimer's disease, amyloid precursor protein APP	APP is a cell surface receptor on neurons that is degraded to amyloid-β peptides Aβ(40–42). Extracellular aggregation of the peptides leads to amyloid plaques that can cause synaptic disconnection. Also, intracellular accumulation of phosphorylated tau protein leads to paired helical filaments and neurofibrillary tangles.
cataracts, crystallins	Protein aggregates in the lens, but unclear if these are amyloids. Proteins have aged (deamidation, disulfide-bond formation).
TTR amyloidosis Type I, transthyretin	Carrier for thyroid hormones and retinol. Homotetramer. > 80 mutations described, most of which are related to amyloid deposition.
Huntington's disease, huntingtin	Possible role in microtubule-mediated transport or vesicle function. Genetic CAG-expansion leads to poly-Gln huntingtin. The Gln-rich sequences can aggregate to infinite β-sheets.

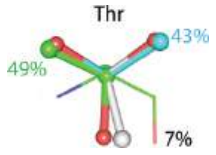
QUESTIONS

18.1 Calculate the concentration of neutral Gly in a 1 M solution at pH = pI. The pK_A values for the carboxyl and ammonium groups are pK_{A1} = 2.3 and pK_{A2} = 9.8, respectively. Is it a good idea to draw the neutral form of free amino acids?

18.2 Calculate the expected percentage of prolyl *cis*-peptide bonds (ΔG = –2 kJ mol^{-1}) and non-prolyl *cis*-peptide bonds (ΔG = –10 kJ mol^{-1}) at 25°C.

18.3 How many chiral centers are there in (a) glycine, (b) isoleucine, (c) tryptophan and (d) threonine?

18.4 The three most prominent rotamers of Thr are shown. The rotamer in cyan accounts for 43% in Thr, but the same conformation accounts for 73% of all rotamers in Val. What could be the explanation for this high propensity?



18.5 The structure of a fourteen-residue peptide displays ψ and ϕ–angles of –60±10°. What can you say about this structure?

18.6 The coordinates of a helix are given by

$$\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} = \begin{pmatrix} R \cdot \cos\dfrac{2\pi n}{r} \\ R \cdot \sin\dfrac{2\pi n}{r} \\ \dfrac{P}{r} \cdot n \end{pmatrix}$$

where n is the residue number, R is the helix radius, r is the number of residues per turn, and P is the pitch. Calculate the coordinates of four consecutive Cα atoms of an ideal α-helix. Show that three consecutive Cα atoms enclose an angle of 90° and the torsion angle between four Cα atoms is 50°.

18.7 Compare the hydrolysis of the phosphoserine in phosphatases with the regeneration of FGly from sulfo-FGly in sulfatases.

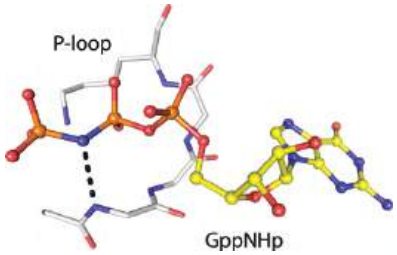
18.8 Protonated 1,3-dicarbonic acids and also 3-ketoacids are unstable. They lose a molecule of CO_2 by a pericyclic reaction. Draw the movement of electrons and the intermediates of the decarboxylation reaction. How is this decarboxylation avoided *in vivo*?

18.9 Draw the two alternative Greek key motifs.

18.10 The monomer-dimer equilibrium $2M \rightleftarrows D$ is characterized by the equilibrium constant K_a. The total monomer concentration is M_0. Derive a formula that relates the mole fraction M/M_0 of monomer M to the equilibrium constant K_a. Draw a graph M/M_0 as a function of M_0 with logarithmic scale for M_0 and K_a values of 1000 M^{-1} and 10^6 M^{-1}. Show that the mole fraction is 0.5 for $M_0 = 1/K_a$ ($= K_d$). Using l'Hôpital's rule, show that the mole fraction approaches unity (complete complex dissociation) for $M_0 \rightarrow 0$ and approaches zero (100% complex) for $M_0 \rightarrow \infty$.

18.11 In hydrophobic environments, a set of five water molecules often forms a planar ring. What could be the explanation? The H-O-H angle in water is 104.5°.

18.12 GppNHp is a non-hydrolyzable GTP analog where the oxygen atom bridging the β- and γ-phosphate is replaced by an NH group. In the figure below, a hydrogen bond is drawn between a peptide NH group of the P-loop and the N-atom of GppNHp. Why is this allowed?



18.13 Using the scheme for alternative ligand binding by induced fit and conformational selection (Figure 18.43), express the binding constant K_{bind} by the other three equilibrium binding constants. Why is K_{select} much larger than $K_{collision}$? K_{conf} is dominated by thermal energy, which interconverts the conformations. This energy is then kept upon ligand binding. Would you expect K_{conf} to be larger or smaller than unity?

18.14 When some Gram-negative bacteria, including *Vibrio* and *Pseudomonas*, encounter high temperature stress, they respond by fast *cis* to *trans* isomerization of unsaturated fatty acids in their membrane lipids. What is the molecular effect and why is it beneficial under this condition?

18.15 When an amphiphilic molecule is mixed with a little water and a large excess of oil, another type of micelle, the *inverted micelle*, is formed. What would be the general structure of such a micelle and what are the driving forces for its formation?

18.16 Formulate the law of mass action for micelle formation of a charged amphiphile, *e.g.* SDS, taking other ions present in the solution and a non-neutral resulting micelle into account.

18.17 Show that the same expression for the total concentration $[S]_{tot,equiv}$ at the equivalence $[S] = n \cdot [M_n]$ is obtained using either variant of the mass conservation:

$$[S]_{tot} = [S] + n \cdot K \cdot [S]^n$$

in terms of monomeric amphiphile or

$$[S]_{tot} = \sqrt[n]{\frac{n \cdot [M_n]}{n \cdot K}} + n \cdot [M_n]$$

in terms of micelle.

18.18 Calculate the *cmc* by using the condition that the third derivative of the equation

$$[S]_{tot} = [S] + n \cdot K \cdot [S]^n$$

is zero:

$$\frac{d^3[S]}{d([S]_{tot})^3} = 0$$

18.19 Derive a formula that allows estimation of the aggregation number n as a function of the measurable quantity $d[S]/d[S]_{tot}$, starting by linearizing the equation.

$$\frac{d[S]}{d[S]_{tot}} = \frac{1}{1 + n^2 \cdot K \cdot [S]^{n-1}}$$

18.20 Show that for large aggregation numbers the approximation $\Delta G_m \approx RT \cdot \ln cmc$ holds.

18.21 The peptide with sequence LRKLRKRLLRLWKLRKR can penetrate membranes. What might be the reason for this property?

18.22 Calculate the minimum number of residues for a straight, non-bent α-helix to perpendicularly traverse a membrane of 3 nm mean thickness. What about a helix tilted by 45°?

18.23 Why can a single β-strand not exist in a membrane? What would be the shortest β-barrel to traverse a membrane?

18.24 Why is guanidinium chloride less destabilizing than guanidinium thiocyanate?

18.25 Guanidinium sulfate stabilizes proteins, what does this tell us about the relative effect of anions *versus* cations in the Hofmeister series?

18.26 If RNase A is (1) denatured and reduced, (2) re-oxidized by air in the presence of denaturant, and (3) dialyzed into native conditions, the resulting scrambled RNase A has 1% activity compared to the starting material. What could be the explanation? Remember that native RNase A has eight Cys tied in four disulfide bonds.

18.27 Why is it not always advisable to purify proteins from thermophilic microorganisms at 4°C?

18.28 The stability of a protein is given by

$$\Delta G_{\text{stab}}(T) = -RT\ln K_{\text{unfold}} = \Delta H_{\text{m}}\left(1-\frac{T}{T_{\text{m}}}\right)+\Delta C_{\text{p}}\left(T-T_{\text{m}}-T\cdot\ln\frac{T}{T_{\text{m}}}\right)$$

Using analytical calculus, show that the temperature of maximum stability is at

$$T = T_{\text{m}}\cdot\exp\left(-\frac{\Delta H_{\text{m}}}{T_{\text{m}}\cdot\Delta C_{\text{p}}}\right)$$

REFERENCES

Basse, N., Kaar, J. L., Settanni, G., Joerger, A. C., Rutherford, T. J. and Fersht, A. R. (2010) Toward the rational design of p53-stabilizing drugs: Probing the surface of the oncogenic Y220C mutant. *Chem Biol* **17**(1): 46–56.
· a destabilized, cancer-causing variant of p53 was stabilized by a small molecule – Stabilization of destabilized proteins by small molecules could be a way to treat disease in general

Berkholz, D. S., Driggers, C. M., Shapovalov, M. V., Dunbrack, R. L., Jr. and Karplus, P. A. (2012) Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc Natl Acad Sci U S A* **109**(2): 449–453.
· statistical analysis of distorted peptide bonds in protein structures

Bhaskara, R. M. and Srinivasan, N. (2011) Stability of domain structures in multi-domain proteins. *Sci Rep* **1**: 40.
· linkers in proteins

Chakrabarti, P. and Janin, J. (2002) Dissecting protein-protein recognition sites. *Proteins* **47**(3): 334–343.
· review on protein-protein recognition sites

Chothia, C. and Janin, J. (1975) Principles of protein-protein recognition. *Nature* **256**(5520): 705–708.
· early review on protein-protein interactions

Chothia, C. (1984) Principles that determine the structure of proteins. *Annu Rev Biochem* **53**: 537–572.
· early review on protein structure

Chou, P. Y. and Fasman, G. D. (1977) β-turns in proteins. *J Mol Biol* **115**(2): 135–175.
· survey of turns in protein structures

Copley, R. R. and Bork, P. (2000) Homology among $(\beta\alpha)_8$ barrels: Implications for the evolution of metabolic pathways. *J Mol Biol* **303**(4): 627–641.
· review on TIM-barrels

Creighton, T. E. (1992) *Proteins: Structures and Molecular Properties*. New York, Freeman.
· a classic book on protein structure and interactions

Creighton, T. E., Ed. (1992) *Protein Folding*. New York, Freeman.

Crick, F. H. C. (1953) The packing of α-Helices: Simple coiled-coils. *Acta Cryst* **6**: 689–697.
· first description of the packing of α-helices – This work helped Crick later in the determination of the helical DNA structure

Davis, B. D. (1958) On the importance of being ionized. *Arch Biochem Biophys* **78**(2): 497–509.
· a classic read on compartmentalization of metabolites in cells

Dunitz, J. D. (1994) The entropic cost of bound water in crystals and biomolecules. *Science* **264**(5159): 670.
· entropy of bound water molecules

Eisenberg, D., Weiss, R. M., Terwilliger, T. C. and Wilcox, W. (1982) Hydrophobic moments and protein structure. *Faraday Symposia of the Chemical Society* **17**: 109–120.

Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**(1): 105–132.
· two hydropathy scales

Go, Y. M. and Jones, D. P. (2008) Redox compartmentalization in eukaryotic cells. *Biochim Biophys Acta* **1780**(11): 1273–1290.
· review on disulfide formation *in vivo*

Guiley, K. Z. and Shokat, K. M. (2023) A small molecule reacts with the p53 somatic mutant Y220C to rescue wild-type thermal stability. *Cancer Discov* **13**(1): 56–69.
· attempt to stabilize p3 carrying the Y220C mutation by a covalent inhibitor

Hartl, F. U. (2017) Protein misfolding diseases. *Ann Rev Biochem* **86**(1): 21–26.
· review on protein folding diseases

Jaenicke, R. (1987) Folding and association of proteins. *Prog Biophys Mol Biol* **49**(2–3): 117–237.
· folding of oligomers

Jaenicke, R. and Böhm, G. (1998) The stability of proteins in extreme environments. *Curr Opin Struct Biol* **8**(6): 738–748.
· review on the stability of proteins from extremophiles

Lawrence, M. C. and Colman, P. M. (1993) Shape complementarity at protein/protein interfaces. *J Mol Biol* **234**(4): 946–950.
· introduces the surface complementarity coefficient

Liu, C. C. and Schultz, P. G. (2010) Adding new chemistries to the genetic code. *Annu Rev Biochem* **79**: 413–444.
· review on the introduction of modified amino acids in proteins

Lupas, A. N. and Gruber, M. (2005) The structure of α-helical coiled coils. *Adv Protein Chem* **70**: 37–78.
· review on α-helical coiled-coils

Matthews, B. W. and Liu, L. (2009) A review about nothing: Are apolar cavities in proteins really empty? *Protein Science: A Publication of the Protein Society* **18**(3): 494–502.
· review on non-polar cavities in proteins

Nödling, A. R., Spear, L. A., Williams, T. L., Luk, L. Y. P. and Tsai, Y. H. (2019) Using genetically incorporated unnatural amino acids to control protein functions in mammalian cells. *Essays Biochem* **63**(2): 237–266.
· Review of non-natural amino acids that have been incorporated in mammalian cells

Olesen, N. E., Westh, P. and Holm, R. (2015) Determination of thermodynamic potentials and the aggregation number for micelles with the mass-action model by isothermal titration calorimetry: A case study on bile salts. *J Colloid Interface Sci* **453**: 79–89.
· mathematical description of the basics of micelle formation

Pace, C. N. (1995) Evaluating contribution of hydrogen bonding and hydrophobic bonding to protein folding. *Meth Enzymol* **259**: 538–554.
· contribution of hydrogen bonds to protein folding

Pain, R. H., Ed. (1994) *Mechanisms of Protein Folding*. Oxford, IRL Press.
· comprehensive references on all aspects of protein folding

Richardson, J. S. and Richardson, D. C. (1988) Amino acid preferences for specific locations at the ends of α-helices. *Science* **240**(4859): 1648–1652.
· review on helix capping

Santoro, M. M. and Bolen, D. W. (1988) Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry* **27**(21): 8063–8068.
· first mathematical description of protein folding and unfolding

Schmid, F. X. (2001) Prolyl isomerases. *Adv Protein Chem* **59**: 243–282.
· review on prolyl isomerases

Sibanda, B. L. and Thornton, J. M. (1985) β-hairpin families in globular proteins. *Nature* **316**(6024): 170–174.
· review on β-hairpins in proteins

Tang, H., Zhang, P. and Luo, X. (2022) Recent technologies for genetic code expansion and their implications on synthetic biology applications. *J Mol Biol* **434**(8): 167382.
· review on non-natural amino acids

Wright, P. E. and Dyson, H. J. (1999) Intrinsically unstructured proteins: Re-assessing the proteissssn structure-function paradigm. *J Mol Biol* **293**(2): 321–331.
· review on intrinsically unfolded proteins

Young, D. D. and Schultz, P. G. (2018) Playing with the molecules of life. *ACS Chem Biol* **13**(4): 854–870.
· review on incorporating non-natural amino acids into proteins by matched pairs of tRNAs and aminoacyl tRNA synthetases

ONLINE RESOURCES

https://proteininformationresource.org/resid
· database of post-translational modifications

heller.userweb.mwn.de/membrane/membrane.html
· structural models of membranes

https://web.iitm.ac.in/bioinfo2/prothermdb
· protherm database

19 Nucleic acids

Nucleic acids are named after their predominant localization in the nucleus (for eukaryotes) and their acidic character. They come in two types, *ribonucleic acid* (RNA) and *deoxy-ribonucleic acid* (DNA). The main function of DNA is to store genetic information, although many viruses use RNA for this purpose. RNA has a wider range of functions. Messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) take over key functions during gene expression and protein biosynthesis. *Ribozymes* are enzymatically active RNA molecules, probably the first enzymes on the planet that existed billions of years ago before any proteins were around. RNAs often exert their structural and catalytic function in conjunction with proteins in *ribonucleoprotein* complexes such as ribosomes, spliceosomes, and telomerases. In addition, many non-coding or *regulatory RNAs* have been discovered over the last decade that regulate genome organization, gene expression, and serve in the innate immune system of bacteria (CRISPR-CAS). Overall, RNA is much more versatile than DNA, both structurally and functionally. By contrast, DNA is chemically more stable than RNA.

Nucleic acids are polymers of nucleotides or in short *polynucleotides*. About ⅔ of their mass is the ribose-phosphate backbone, whereas the rest is accounted for by the nucleobases (Figure 19.1). Nucleic acids show excellent solubility in comparison to proteins because of their charged ribose-phosphate backbone; the hydrophobic areas of the nucleobases are buried (Table 18.1). Similar to amino acids in proteins, it is the nucleotide sequence that determines the information content, the final structure, and the function of nucleic acids. We will begin this Chapter with a chemical and geometric view on the constituents of nucleic acids (Sections 19.1–19.3), and then look at their assembly into secondary and tertiary structures (Sections 19.4–19.6).

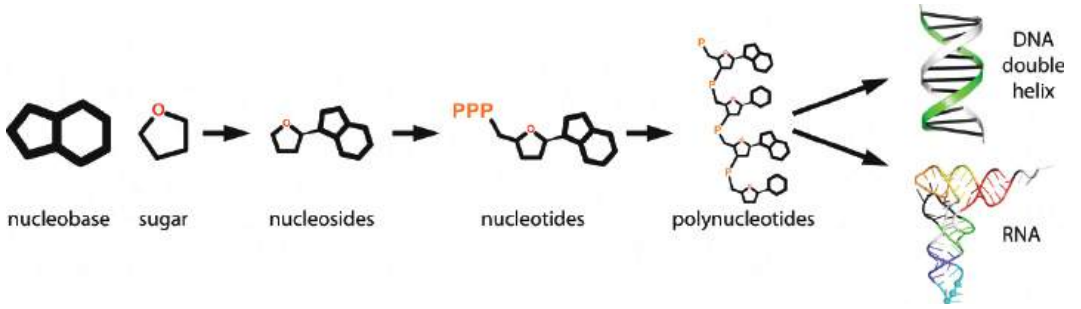


FIGURE 19.1: From simple organic groups to nucleic acids. Nucleobases combine with sugars to nucleosides, which are phosphorylated to nucleotides and then polymerized into polynucleotides. Two strands of a polynucleotide can form a double helix or, in the case of RNA, a single strand can fold into a complex three-dimensional shape.

19.1 NUCLEOBASES, NUCLEOSIDES, AND NUCLEOTIDES

The number of canonic building blocks that make up nucleic acids is smaller than for proteins: just four nucleobases compared to the twenty standard amino acids. A nucleobase is a flat, aromatic heterocycle, some of which have weakly basic character, hence the name nucleo*base*. The hydrophobic surfaces of nucleobases engage in non-polar interactions whereas the nitrogen and oxygen atoms can entertain polar interactions, predominantly hydrogen bonds. The four nucleobases found in DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). RNA contains uracil (U) instead of T (Figure 19.2). U and T are identical but for the extra methyl group in T. A and G consist of a fused ring system derived from purine, C and T/U are pyrimidine derivatives. The pK_A values of protonated G and C are in the range of carbonic acids (acetic acid has pK_A = 4.7). Protonated G and C are not observed in solution but do occur in base triplets that are shielded from solvent (Section 19.5.2).

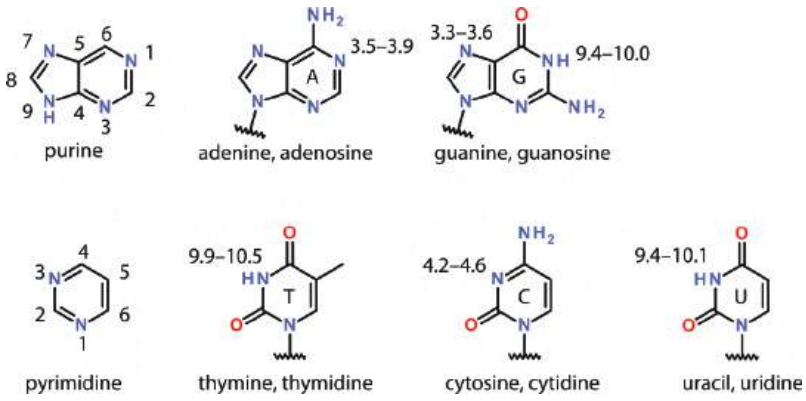


FIGURE 19.2: Structure and numbering of nucleobases. Adenine and guanine are derivatives of purine. Using the purine numbering system, adenine is 6-amino purine. Thymine, cytosine, and uracil are derivatives of pyrimidine. Note that uracil is thymine that lacks the methyl group at the 5-position. The most stable tautomers of the bases are shown. The wavy line indicates the glycosidic bond to the (deoxy)ribose in the nucleosides adenosine, guanosine, thymidine, cytidine, and uridine. Number ranges next to the nitrogen atoms in rings are pK_A values found in nucleosides and nucleotides.

The unit comprised of a nucleobase and a (deoxy)ribose is termed *nucleoside*. In purines it is the N9-atom and in pyrimidines the N1-atom that forms a *glycosidic bond* with the C1'-atom of a ribose sugar (Figure 19.3). *Nucleotides* are phosphorylated nucleosides. For instance, ATP is adenosine 5'-triphosphate, an adenosine nucleoside that has a string of three phosphoryl groups attached to the 5'-hydroxyl group of ribose. Figure 19.3 illustrates the blueprint of nucleosides and nucleotides with

the nucleoside cytidine (no phosphoryl group) and the nucleotides dGTP and cAMP as examples. The nucleotides dGTP, dCTP, dTTP, and dATP with 2'-deoxyribose as the sugar are building blocks of DNA. In contrast, ribose is the sugar in the RNA-forming nucleotides GTP, CTP, UTP, and ATP. In cAMP, an ATP derivative that functions as second messenger in signal transduction, the phosphoryl group, forms a cyclic phosphodiester with two hydroxyl groups of ribose.

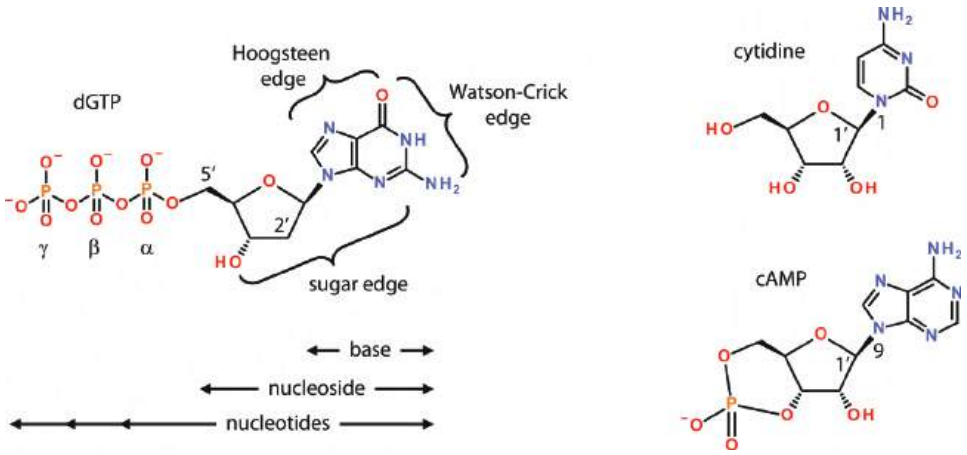


FIGURE 19.3: Nucleobases, nucleosides, and nucleotides. The ribose condenses with a nucleobase to form a nucleoside. A nucleotide is a phosphorylated nucleoside. dGTP is 2'-deoxy guanosine 5'-triphosphate and cAMP is cyclic 2',5'-adenosine monophosphate. For the numbering of the ribose see Section 19.2. Phosphate groups are labeled with Greek letters. The Watson-Crick edge contains the hydrogen bond donors and acceptors in nucleobases that form standard base pairs. The Hoogsteen and sugar edges are used in non-standard base pairs and are involved in tertiary interactions, especially in RNA.

The pK_A values of phosphoric acid are 2.0, 6.8, and 12.5. The acidity of the phosphate groups changes slightly in ATP, which has pK_A values of 0.9, 1.5, 2.3, and 7.7. The acidity of the phosphates in all nucleotides is higher than the basicity of the nucleobases, so all nucleotides are strong acids and are negatively charged under physiological conditions. Nucleotide triphosphates have 3–4 negative charges that can engage in polar interactions, and strongly attract cations. Most nucleotides and also nucleic acids are bound to metal ions (Mg^{2+}, Na^{+}) in solution.

19.1.1 Hydrogen bonds between nucleobases – base pairs

The *Watson-Crick* edge in nucleobases is the side that forms hydrogen bonds in the standard Watson-Crick base pairs, which are the dominant pairings in all double-helical nucleic acid structures (Figure 19.4). In the beginning it was unclear how the nucleobases might interact with each other. A key piece of information came with the analysis of DNA composition in the 1940s by Erwin Chargaff, who showed that the molar ratio of adenosine and thymidine in double-stranded DNA is always unity ([*A*] = [*T*]), and the same is true for the ratio of guanosine and cytidine ([*G*] = [*C*]). These *Chargaff rules* suggested A-T and G-C base pairs. There are two hydrogen bonds in A-T and three in G-C pairs, which makes the G-C pair the more stable one. The hydrogen bonding pattern establishes selectivity for the base pairings.

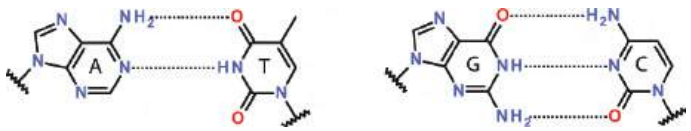


FIGURE 19.4: Standard Watson-Crick base pairs. An A-T base pair forms two hydrogen bonds. The more stable G-C base pair has three hydrogen bonds. The two pairs are isosteric, *i.e.* they have roughly the same overall size, which is important when they stack to form a double helix in DNA and RNA. The biologically relevant tautomers are shown.

The nucleobase structures in Figure 19.2 show the most stable tautomers, the keto- and amino-forms. The less stable enol- and imino-forms are possible in base-pairing but occur rarely. If they do, the tautomerism can lead to mutations during DNA replication. For example, the enol form of adenine can cause A-to-C transversions and the enol form of thymine can cause T-to-A transversions.

Several other nucleobase orientations and combinations also allow formation of hydrogen bonds, among them non-Watson-Crick base pairs, reverse Watson-Crick base pairs, Hoogsteen base pairs, and reverse Hoogsteen base pairs (Figure 19.5). Reverse base pairs are formed between bases in parallel strands with the same polarity, putting the glycosidic bonds on opposite (*reverse*) sides of the base pair.

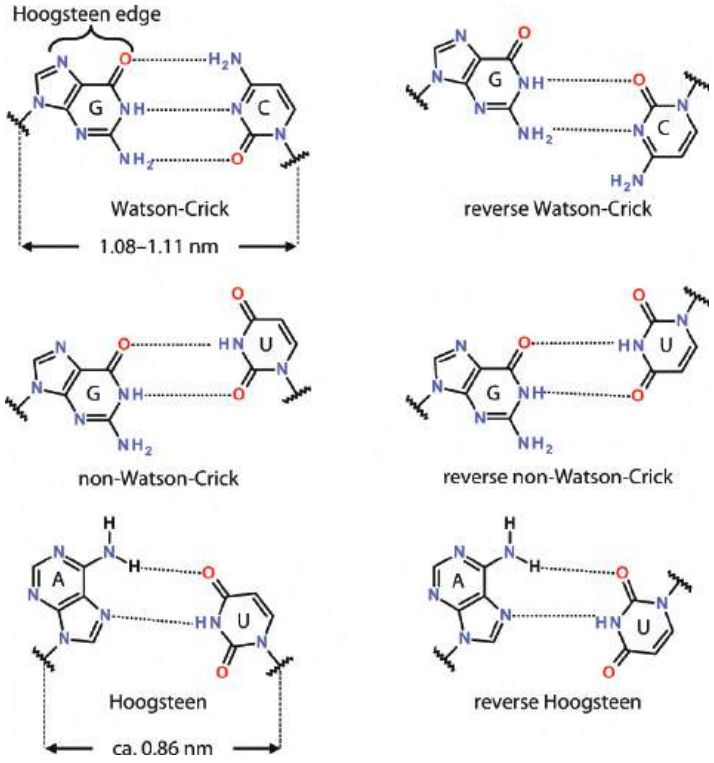


FIGURE 19.5: Non-standard base pairs. Base pairing is dictated by hydrogen bonds between bases with matching donors and acceptors. The distances between the C1′ atoms (wiggly lines) in Watson-Crick base pairs range between 1.08 nm and 1.1 nm. The nucleobases are on the same side in a *cis*-like orientation of their glycosidic bonds, as opposed to reverse base pairs (right-hand column) where the glycosidic bonds are oriented antiparallel. Non-Watson-Crick base pairs such as G-U are frequently observed in tRNA. The purine in Hoogsteen pairs has *syn*-conformation and the C1′-C1′ distance is only 0.86 nm.

Non-Watson-Crick base pairs use the Watson-Crick edge of the nucleobases, but pair with the "wrong" nucleobase. These include the base pairs G-G, G-A, G-T, A-A, A-C, T-T, T-C, C-C, and G-U. Non-Watson-Crick base pairs have fewer or weaker hydrogen bonds than usual base pairs. Their formation can lead to replication errors and mutations if they occur in DNA. Some non-Watson-Crick base pairs are formed between mRNA and tRNAs during translation. The non-Watson-Crick G-U pair is particularly important because it allows the same tRNA to read out two different codons (see Figure 19.35). In G-U pairs the U is shifted in the plane of the base pair to match the hydrogen bond donors and acceptors of G and to form two hydrogen bonds (Figure 19.5).

Hoogsteen pairing always involves the N7 atom of a purine in the *Hoogsteen edge* (Figure 19.5) and the Watson-Crick edge of a pyrimidine. Their formation leads to smaller distances between the C1'-atoms compared to Watson-Crick pairs, which makes them unsuited for regular DNA structures. Instead, Hoogsteen pairs are prominent in RNA tertiary structure, Z-DNA, and DNA structures containing more than two strands (Section 19.5.2). Reverse base pairs, particularly reverse Hoogsteen base pairs, are present in DNA triplexes (Section 19.5.2.1) and frequently observed in folded RNA.

19.1.2 Non-standard nucleobases in DNA

Similar to the post-translational modifications of side chains in proteins, the nucleobases in DNA and RNA are subject to modification, which expands the information content and biophysical properties of nucleic acids. One of the most frequent modifications in DNA is methylation (Figure 19.6), as in 5-methyl-cytosine, 5-hydroxymethyl-cytosine, and N^6-methyl-adenine. Depending on where the methyl group is added, it can add positive charges, change pK_A values and hydrogen-bonding capabilities, and change the stacking properties of the nucleobases. In DNA of higher eukaryotes, the amount of cytosine methylation depends on the organism, its developmental stage, and the tissue. It is estimated that 1.3–1.5% of all cytosine nucleotides in mouse and human brain DNA are methylated. In mammals, the degree of methylation is high during neurogenesis and decreases with age. This additional, sequence-independent level of information encoded in eukaryotic DNA is part of *epigenetics*. The degree of methylation is controlled by DNA methyl transferases. Bacteria use DNA methylation to distinguish their own DNA from foreign DNA, *e.g.* from bacteriophages. Methylation is also used in bacteria to distinguish parent DNA from recently replicated DNA. If a mismatch occurs during DNA replication, the error is repaired taking the sequence of the methylated parent DNA as the reference.



FIGURE 19.6: Modified bases in DNA. 5-methyl-C is by far the most frequently modified nucleobase in DNA, followed by 5-hydroxymethyl-C and N^6-methyl-A. None of these modifications change the base-pairing abilities of the nucleobase.

19.1.3 Non-standard nucleobases in RNA

RNA shows much more variation of nucleobases than DNA due to a large variety of post-transcriptional modifications. The modifications are introduced by dedicated RNA-modifying enzymes, either by editing the nucleobases *in situ* or by replacement of the entire nucleobase. The Modomics database lists over 170 different post-transcriptional RNA modifications on the levels of the nucleobase, nucleoside, and nucleotide. The modifications include differences in bond order, reduction of bonds, exchange of single atoms in nucleobases and phosphates, different sugars, addition of the small methyl group to the nucleobase and sugar, and addition of larger groups to the nucleobase that may even double its molecular mass. The largest number of modifications is present in tRNAs. All tRNAs contain 5-methyl uracil (thymine), pseudouracil, and dihydrouracil (Figure 19.7), which engage in hydrogen bonding and tertiary interactions that stabilize the characteristic L-shape of tRNAs (Section 19.6.2). The saturated C5 and C6 atoms in dihydrouracil lead to non-planarity (puckering) of the pyrimidine ring. Formally, pseudo-uracil is a uracil in which the N1 and C5 atoms are exchanged, leading to an unusual C-glycosidic bond. Thymine is a regular component of DNA, but also occurs in tRNA where it is bound to ribose as ribothymidine. 4-thiouridine has a sulfur atom in place of the usual oxygen. This fluorescent nucleobase happens to be quite reactive and is often used for fluorescent and spin labeling, as well as crosslinking studies. Its biological role seems to be a near-UV (UVA; 310–400 nm) radiation sensor in bacteria and archaea. Irradiation of *E. coli* with UVA induces a growth delay, probably mediated by a crosslink in tRNAs between 4-thiouridine and a nearby cytidine.

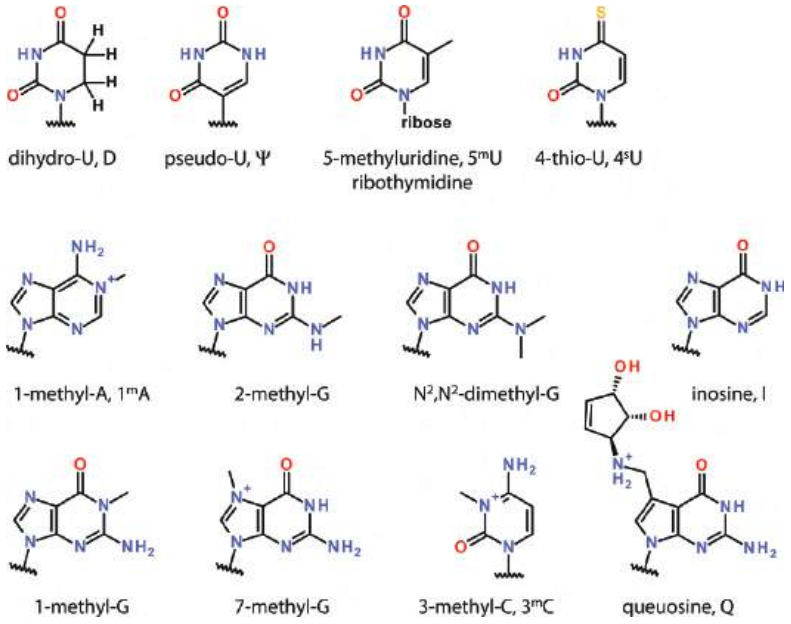


FIGURE 19.7: Modified bases in tRNA. These modified bases are present in tRNA where they stabilize the structure by conserved interactions or help decoding mRNA. Methyl groups are indicated as "Me". If present at the third (*wobble*) position of the anticodon 5′-IAU-3′, inosine can pair with A, U, and C and thus decode the mRNA triplets AUA, AUU, and AUC, which all code for Ile. Queuosine replaces the G at the wobble position of tRNAs, with GUN anticodons coding for Asp, Asn, His, and Tyr.

Methylation is another very common modification in tRNA. Methylation adds a small non-polar group. Depending on the position, methylation can lead to permanent charges and shifts in pK_A values. Adenine can be methylated at N1 and N6 (as in DNA); guanine at N1, N2, and N7; cytosine at N3; and uracil at C5 (Figure 19.7). Methylation of adenosine at position 1 introduces a permanent positive charge under physiological conditions and increases the hydrophobic surface of the nucleobase for stacking interactions (Section 19.4.1). At the same time, the Watson-Crick edge cannot form the standard base pair with uracil due to steric repulsion. In addition to tRNA, ribosomal RNA (rRNA) also contains dozens of modifications (mostly methylation), which cluster at the catalytic sites for peptide bond formation and translocation in the ribosome. Eukaryotic mRNA carries a cap structure at its 5'-end, a methylated G with a 5'–5'-triphosphate that is recognized during the initiation of translation.

Several more radically modified nucleobases are specific for particular tRNAs. For example, queuosine (Q) bases have a pentenyl ring attached to the N7-atom of guanosine *via* a methyl-amino linker. Q bases are present at the third position of the tRNA anticodon where they pair with U and sometimes also with C. The alternative pairings enable the same tRNA to recognize different codons on the mRNA. The fact that there are less tRNAs in a cell (maximum of 41) than mRNA codons for amino acids (61) has been termed the *wobble hypothesis*. We will return to this hypothesis when we look at the various base-pairing possibilities in Section 19.4.2.

19.2 RIBOSE AND NUCLEOBASE CONFORMATIONS

The nucleobase in nucleotides is bound to a sugar by an N-glycosidic bond. The sugar has a 5-membered ring that resembles the heterocycle tetrahydrofuran, hence such sugars are called *furanoses*. Ribose in RNA and 2'-deoxyribose in DNA are furanoses that differ only by the presence or absence of a hydroxyl group at the second carbon atom (Figure 19.8). To distinguish the atoms of the nucleobases from those of the ribose, a prime (') is used for the sugar numbering.

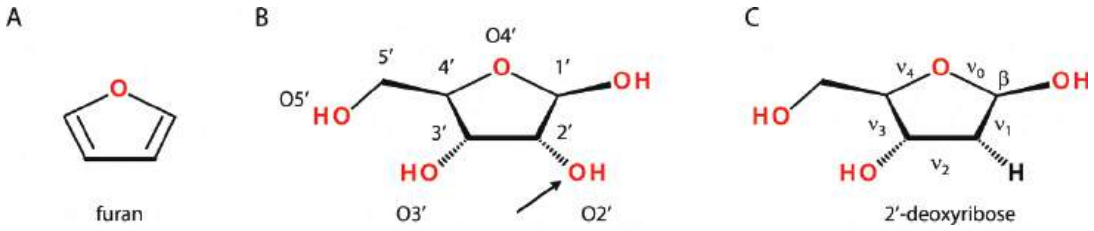


FIGURE 19.8: Ribose nomenclature. A: The saturated furan derivative tetrahydrofuran makes up the framework for furanoses. B: Ribose and 2'-deoxyribose differ only by the hydroxyl group at C2' (arrow). C: Nucleic acids exclusively contain the diastereomer shown, the β-anomer. The torsion angles within the ribose ring are labeled ν_0 to ν_4.

The C1' atom of ribose is the *anomeric* carbon, which gives rise to two diastereomers termed *anomers* (Section 17.1): α-anomers have the 1'-hydroxyl group below the plane of the furanose ring (as drawn in Figure 19.8), in the β-anomer it is above the plane. While the two anomers can interconvert in the free sugar, in nucleosides, nucleotides, and nucleic acids the β-anomeric configuration is fixed. The conformation of sugars is described by torsion angles ν_0 to ν_4. In riboses, these angles are not as tightly restricted as the peptide backbone angles, but they are still highly correlated: changing one torsion angle automatically induces concomitant rotations about the other four bonds.

19.2.1 SUGAR PUCKER

In a planar conformation, the 5-membered ring of ribose and deoxyribose would have all carbon atoms in the eclipsed conformation, which corresponds to maximal torsional energy (Section 17.2.1.2). The most energetically favorable conformations of the ribose are therefore non-planar;

the degree of non-planarity is termed *puckering*. Puckering is a general way to reduce the torsional energy in a ring. The most stable conformations of the sugar are termed *envelope* and *twist*. In the envelope conformation, four ring atoms are almost co-planar and the fifth atom is above (*endo*) or below (*exo*) this plane. Nucleobases in nucleosides and nucleotides always lie in the *endo* face of the ribose. The twist conformation has one atom above and one below the plane defined by atoms C1', C4', and O4' (Figure 19.9). Twist conformations are somewhat less stable than the envelope conformation and sometimes occur in tRNA.

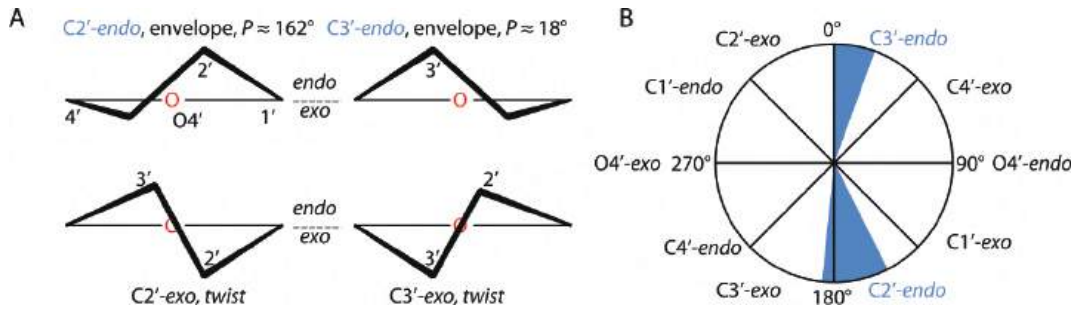


FIGURE 19.9: Ribose conformations. A: The conformations are defined relative to the plane that is formed by the atoms C1', C4', and O4'. The view is side-on to the plane so that the three atoms appear as a line. The conformations are drawn in an exaggerated fashion to highlight their differences. Envelope conformations have four atoms in a plane while in *exo*-conformations only three out of the five ribose atoms are in plane. B: The circular plot shows how the phase angle of the ribose ring corresponds to the various *endo*- and *exo*-conformations. The energetically preferred C2'-*endo* and C3'-*endo* regions are marked in blue.

The overall conformation of the sugar can be described by the sugar pucker, a convenient way to quantify the degree of sugar non-planarity in nucleic acids. The sugar pucker is defined by the five torsion angles ν_0 to ν_4 of the ribose ring (Figure 19.8):

$$\text{pucker} = \frac{\nu_2}{\cos P} \qquad \text{eq. 19.1}$$

with the *pseudo-rotation or phase angle P* that depends on the five individual torsion angles as

$$\tan P = \frac{(\nu_4 + \nu_1) - (\nu_3 + \nu_0)}{2\nu_2(\sin 36° + \sin 72°)} \qquad \text{eq. 19.2}$$

The most important sugar pucker in nucleic acids is the deviation of the C2' and C3' atoms from a plane defined by the other three atoms. C2'-*endo* conformations have phase angles between 144° and 190° and are characteristic for B-DNA (Section 19.5.1). Angles between zero and 18° define C3'-*endo* conformations, which are found in A-DNA and helical RNA structures (Section 19.6). These phase angles correspond to sugar puckers in the 25°–45° range. C2'-*endo* and C3'-*endo* are both low-energy conformations of a ribose with little effect on the orientation of the nucleobase. In the context of double-stranded nucleic acids, however, the conformations have different effects (Section 19.5.1).

19.2.2 Syn- and anti- conformations

The orientation of the nucleobase with respect to the ribose is a major energetic factor for nucleotide conformation. This orientation depends on the torsion angle χ, which defines the rotation about the glycosidic bond. In the overwhelming majority of all nucleobases, χ is in the ranges 90° to 180° and −90° to −180°, the *anti*-conformation. The energetically most favored range is $-180° < \chi < -115°$ for pyrimidines and $-180 < \chi < -60°$ for purines. χ angles between −90° and 90° are characteristic of the *syn*-conformation, with the nucleobase above the ring (Figure 19.10).

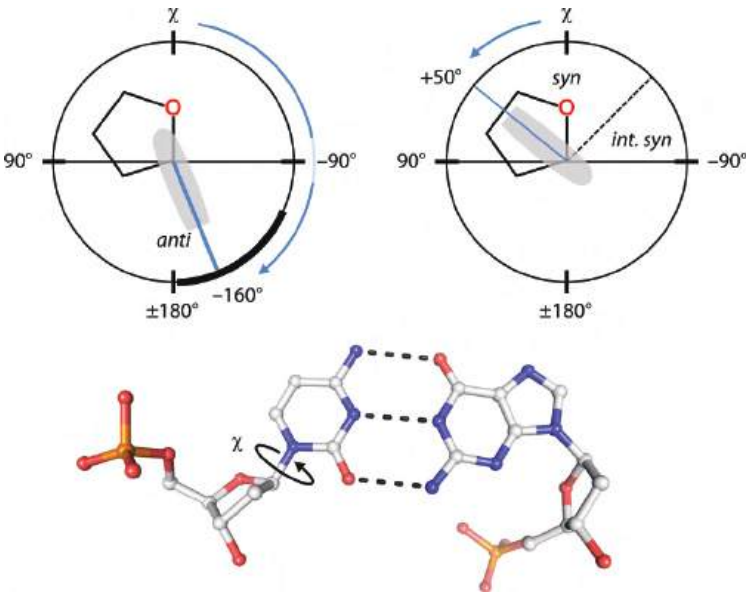


FIGURE 19.10: Nucleobase conformations. The ribose and nucleobase are depicted as a tetrahydrofuran and a gray rod, respectively. Rotation about the glycosidic bond can place the nucleobase away from the ribose (*anti*) or on top of it (*syn*). Intermediate *syn*-conformations are the sub-category with $-45° > \chi \geq -90°$. The lower part of the figure shows a CG base pair in Z-DNA conformation. Cytosine is *anti* and guanine is *syn*.

In the *anti*-conformation, there are only minimal clashes between atoms of the nucleobase and the ribose. The *syn*-conformation will inevitably place atoms of the nucleobase close to the ribose. Therefore, *syn*-conformations are of higher energy than *anti*-conformations and are much less prevalent than the *anti*-conformations. The clashes between nucleobase and ribose are not too severe for purine nucleotides because of their relatively small 5-membered ring above the ribose. Pyrimidine nucleotides with their larger six-membered ring clash more severely with the ribose in the *syn*-conformation, especially the exocyclic carbonyl oxygen at C4. For these reasons, *syn*-conformations are four times more frequent in purine nucleotides than in pyrimidine nucleotides. Many *syn*-nucleotides participate in tertiary interactions, interact with ligands, or contribute to ribozyme active sites. The *syn*/*anti*-conformation may even define the overall conformation of DNA: a stretch of guanine nucleotides in the *syn*-conformation can be placed into a left-handed polynucleotide helix, the Z-DNA (Figure 19.17).

19.3 PRIMARY STRUCTURE OF NUCLEIC ACIDS

Nucleotides are joined together to form *single-stranded* nucleic acids by phosphodiester bonds that connect the 3'- and 5'-hydroxyl groups of neighboring nucleotides. The *polarity* of the strand is given in the 5'→3'-direction. The sequence of nucleobases defines the primary structure of a nucleic acid, just like the sequence of amino acids defines the primary structure of proteins. The only differences between DNA and RNA are the nucleobase thymine (T) and 2'-deoxyribose in DNA compared to the nucleobase uridine (U) and ribose in RNA.

19.3.1 Nucleic acid sequences and codon usage

Similar to proteins, nucleic acid sequences can be compared with databases to identify similar sequences of known function (a *homolog*). Sequences with the same function but from different species are called *orthologs*. *Paralogs* are sequences from the same organism, generated by duplication and variation, leading to new functions of the copy. Nucleic acid databases exist for different organisms. DNA databases may be subdivided into coding and non-coding regions. Specific RNA

databases exist for rRNA, tRNA, or plant RNA. Nucleic acid sequences contain many functional signatures such as splice sites (RNA), restriction sites (DNA), and binding sequences for transcription factors and other proteins (both). Non-coding RNA sequences, which surpass the number of protein-coding RNA sequences, can be analyzed for inverted repeats, sequences involved in transcriptional regulation, and presence of riboswitches (Section 19.6.2).

For a protein-coding gene or spliced mRNA sequence, the *codon usage* is an important aspect. We have seen in Section 18.1.5 that the genetic code is degenerate, and many amino acids are encoded by up to 6 synonymous codons. The frequency of each synonymous codon roughly matches the relative concentration of the decoding tRNA, which ensures efficient translation of an mRNA. The frequency differs between species and the same mRNA sequence may therefore be translated at different velocities in different species. When a protein from one organism (*e.g.* human) is produced recombinantly in a different host (*e.g. E. coli*), the gene sequence is often optimized to match the codon preferences of the host organism to maximize translation efficiency. Codon-optimized sequences can be generated using online servers. Table 19.1 compares the codon usage of human and *E. coli* genes.

TABLE 19.1:
Codon usage (in %) in *E. coli* and *H. sapiens*.

Codon	a.a.	*E. coli*/ human	Codon	a.a.	*E. coli*/ human	Codon	a.a.	57/44	Codon	a.a	*E. coli*/ human
UUU	F	58/46	UCU	S	14/19	UAU	Y	57/44	UGU	C	45/46
UUC	F	42/54	UCC	S	15/22	UAC	Y	43/56	UGC	C	55/54
UUA	L	13/8	UCA	S	14/15	UAA	*	59/30	UGA	*	33/47
UUG	L	13/13	UCG	S	15/5	UAG	*	8/23	UGG	W	100/100
CUU	L	11/13	CCU	P	17/29	CAU	H	58/42	CGU	R	36/8
CUC	L	10/20	CCC	P	13/32	CAC	H	42/58	CGC	R	37/18
CUA	L	4/7	CCA	P	19/28	CAA	Q	33/27	CGA	R	7/11
CUG	L	49/40	CCG	P	51/11	CAG	Q	67/73	CGG	R	11/20
AUU	I	50/36	ACU	T	17/25	AAU	N	47/47	AGU	S	16/15
AUC	I	40/47	ACC	T	41/36	AAC	N	53/53	AGC	S	27/24
AUA	I	9/17	ACA	T	15/28	AAA	K	76/43	AGA	R	5/21
AUG	M	100/100	ACG	T	27/11	AAG	K	24/57	AGG	R	3/21
GUU	V	26/18	GCU	A	16/27	GAU	D	63/46	GGU	G	33/16
GUC	V	21/24	GCC	A	27/40	GAC	D	37/54	GGC	G	39/34
GUA	V	16/12	GCA	A	22/23	GAA	E	68/42	GGA	G	12/25
GUG	V	37/46	GCG	A	35/11	GAG	E	32/58	GGG	G	16/25

Data (in %) were generated using the Codon Usage Database. The first number represents *E. coli*, the second the human codon. The coded amino acid is given in single letter code. * denotes a stop codon. Codons with > 50% difference in usage between *E. coli* and human are colored red (more frequent in *E. coli*) or blue (more frequent in *H. sapiens*).

19.3.2 Backbone conformation

In single-stranded nucleic acids, the hydrophobic surfaces of adjacent nucleobases engage in *stacking interactions* (Section 19.4). This manifestation of the hydrophobic effect leads to burial of non-polar surface areas and solvent-exposure of the polar ribose-phosphate backbone, as well as the nitrogen and oxygen atoms of the nucleobases. Base stacking is the driving force for nucleic acid folding. It imposes a helical structure onto the ribose-phosphate backbone; even dinucleotides and short stretches of single-stranded RNA adopt a helical structure. The extent of the aromatic system and hence the nature of the nucleobases determines the stability of the stacking interaction: interactions between two purines are stronger than between a purine and a pyrimidine, which in turn are stronger than stacking between two pyrimidines. We will revisit base-pair stacking in the context of the DNA double helix (Section 19.5.1) and RNA structure (Section 19.6).

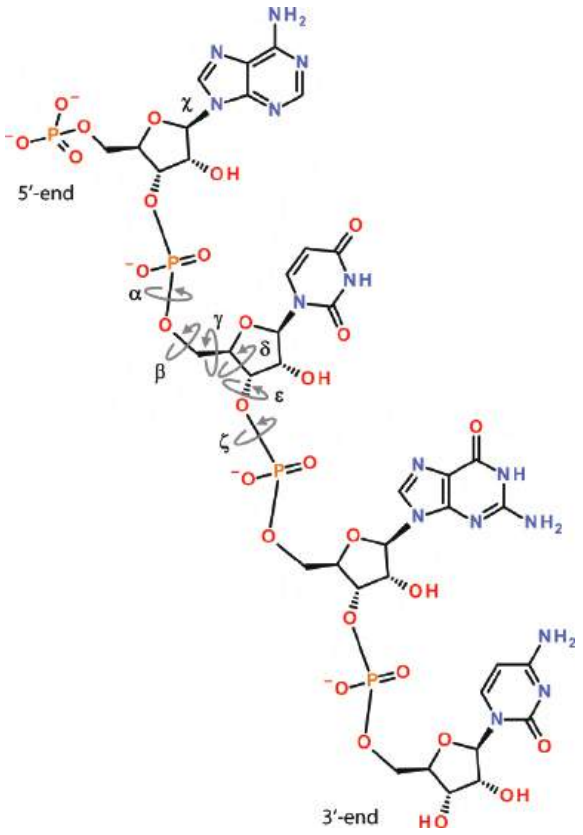


FIGURE 19.11: Backbone structure of nucleic acids. Shown is an RNA tetranucleotide of sequence 5′-AUGC-3′. Sequences are generally written in the 5′→3′-direction. The 5′- and 3′-ends are chemically distinct and establish strand polarity. The six torsion angles along the ribose-phosphate backbone are labeled. Note that the δ-angle is identical to the ν_3-angle in Figure 19.8.

We have seen in Section 18.1.2 that we can describe the conformation around a peptide bond and the resulting energy potential by the torsion angles ϕ and ψ. These torsion angles adopt values that correspond to low energies in the Ramachandran energy surface (Figure 18.21). The equivalent potential for nucleic acids is more complicated because there are eleven rotatable bonds per nucleotide, distributed along the ribose-phosphate backbone and within the ribose (Figure 19.11). Single-stranded nucleic acids are therefore inherently much more flexible than proteins. The torsion angles in nucleic acids adopt values that correspond to energetically favorable conformations with minimized clashes. β-angles, for instance, center around 180±20°, which corresponds to the energetically favored staggered conformation (Figure 17.4) of the hydrogen atoms at C5' and the lone pairs at O5'. Some of the torsion angles are also strongly correlated. α-angles center around 300±20°

and correlate with a narrow γ-angle distribution of around 50±10°. Likewise, the ε-angles are about 190±20° and correlate well with ζ-angles in the 270±20° range. Since the χ-angle of the nucleobase is just opposite to the γ-angle, these two are also highly correlated: the *syn*-conformation correlates with a γ-angle of 180±30°.

The set of torsion angles in the ribose ring can be conveniently summarized by a pseudo-rotation angle *P* (Section 19.2.1 and eq. 19.2) with its prominent energetic minima at the *endo*- and *exo*-conformations. The sugar conformation is correlated with the glycosidic angle χ (*syn*- and *anti*-conformations) of the nucleobases in the double-helical nucleic acid structures (Sections 19.5 and 19.6). Similar to the ϕ/ψ in the Ramachandran plot (Figure 18.21), energy surfaces for the conformation of nucleic acids can be drawn as a function of the various *P*/χ angle combinations (Figure 19.12).

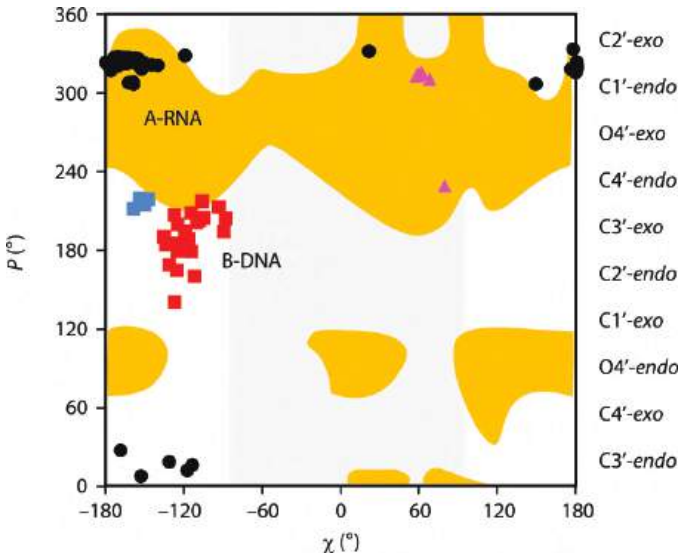


FIGURE 19.12: Schematic energy potential for the pseudo-rotation angle *P versus* the glycosidic angle χ. This kind of representation is analogous to the Ramachandran plot for proteins. The potentials are different for purines and pyrimidines, and also for ribose and 2′-deoxyribose. The yellow areas are the forbidden angle combinations for a 2′-deoxy purine nucleoside where the calculated energy exceeds 24 kJ mol^{-1}. Adapted from Saran *et al.* (1973). Theor. Chim. Acta 30: 31–44. Different ribose conformations are marked on the right-hand side. Note that the C2′-*endo* and C3′-*exo* conformations have similar pseudo-rotation angles of ($P \approx 180 \pm 20°$). The same is true for the C2′-*exo* and C3′-*endo* conformations ($P \approx 0 \pm 20°$). Black circles represent the RNA geometry of a riboswitch with mostly A-RNA geometry (PDB-ID 4rzd). RNA can access conformations that are forbidden for DNA. Red squares represent angles found in regular B-DNA (PDB-ID 1bna). Z-DNA (PDB-ID 4ocb) is represented by pyrimidine nucleotides in *anti*-conformation (blue) and purine nucleotides in *syn*-conformation (magenta). The area shaded in gray is the *syn*-conformational space of $-90° < \chi < 90°$.

In summary, the correlation between the torsion angles in nucleotides leads to energetically preferred conformations of the entire polymer (see Table 19.2). While the low-energy torsion angle combinations ensure a minimum of steric clashes between atoms, most of the stabilization energy in nucleic acid structures comes from hydrogen bonds between bases and the stacking of base pairs, which we will take a look at next.

19.4 BASE-PAIR GEOMETRY AND STACKING

At first glance it seems obvious that base pairs are stabilized by the hydrogen bonds forming between complementary bases. However, under physiological conditions, bulk water competes with the bases for hydrogen bonds: in order to form a single hydrogen bond between nucleobases, two hydrogen bonds with two water molecules need to be broken. The net energy contribution of hydrogen bonds

to base-pair stability may therefore be quite small, similar to what is observed with proteins (Section 17.2.2.3). This leaves base stacking as the main energetic contributor to nucleic acid stability.

19.4.1 Importance of base-pair stacking for double helix formation

The sequence of nucleotides in a single-stranded nucleic acid establishes a unique pattern of hydrogen bond donors and acceptors at the Watson-Crick edges of the nucleobases. Thus, two sequences of complementary nucleobases can associate into a duplex, a spontaneous process that results in a stable double-helical structure. The double helix is further stabilized by van der Waals interactions between base pairs stacked on top of each other. The base stacking in double helices provides a higher energetic contribution than in single-stranded nucleic acids because double helix formation establishes a hydrophobic core in the center, extending along the helical axis, where the non-polar faces of the nucleobases are shielded from solvent. Similar to protein folding, the hydrophobic effect is an important driving force for the assembly of complementary polynucleotide single-strands into a double helix. Duplex formation minimizes solvent-exposed hydrophobic surface areas of the nucleobases. The generation of a hydrophobic environment at the same time increases the strength of the base-pair hydrogen bonds due to the smaller ε_r. The phosphate backbone is on the outside of the helix, which minimizes electrostatic repulsion of the negative charges. In solution the negative charges of the phosphate backbone are partially neutralized by weakly bound cations (Mg^{2+}, Na^{+}). These principles go beyond double-stranded DNA and are also important for triple and quadruple DNA helices, and for folding of RNA molecules. Just like helices in proteins, the helix of nucleic acids is described by general characteristics, including its diameter, axial rise, and pitch (Figure 19.13).

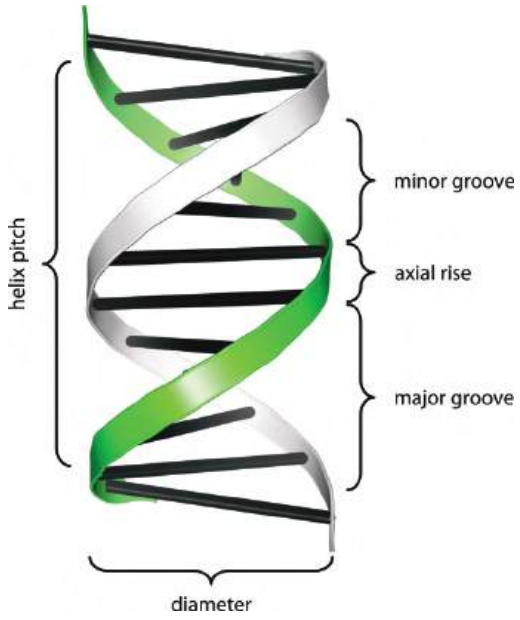


FIGURE 19.13: Schematic depiction of the DNA double helix. The helix is composed of two antiparallel and complementary strands (green and gray with base pairs as black rods) that bury the base pairs in the center and expose the phosphate backbone to solvent. A major and a minor groove run parallel to the strands. Helical parameters including the pitch, the diameter, the number of nucleotides for a full rotation, and the width of the major and minor grooves depend on the actual conformation of the helix (see Table 19.2).

Stacking interactions have been determined experimentally by measuring the effect of changing two bases in the center of an otherwise identical duplex. Duplex stabilities can be measured through changes in absorption; such measurements provide the melting temperature T_m (Section 21.2.7.4; Box 21.5). The stacking energies determined for dinucleotides were then used to

calculate the stabilization by stacking for other sequences. Interestingly, these energies correlated well with the experimentally determined T_m values for large double-stranded B-DNA (Figure 19.14). The correlation indicates that stacking of adjacent nucleobases is the major determinant for double-helix stability, whereas hydrogen bonding provides only a minor contribution.

AT-rich sequences have the lowest stacking energy. As a consequence, an AT-rich duplex (or AU-rich for RNA) melts at lower T_m compared to GC-rich duplexes. However, the stability of a double-stranded DNA depends not only on the composition (type of nucleobase) but also on the sequence of the base pairs (Figure 19.14). The most stable stacking interaction is between a C-G pair and an adjacent G-C pair. If the order is changed to G-C packing on C-G, the stacking energy is almost halved, which is explained by the larger overlap of the nucleobases in a 3'→5' purine-pyrimidine stack compared to a pyrimidine-purine stack. Similarly, if two G-C pairs stack onto each other, the stacking interaction is weaker than for the G-C and C-G pairs. On the other end of the energy scale, the trend is the same for A-T base pairs. A repetitive 5'-AT-3' sequence may form a "hinge" site in DNA: the weak stacking energy of this sequence provides more flexibility to the duplex, enabling it to kink at such sequences (Section 19.5.3.3).

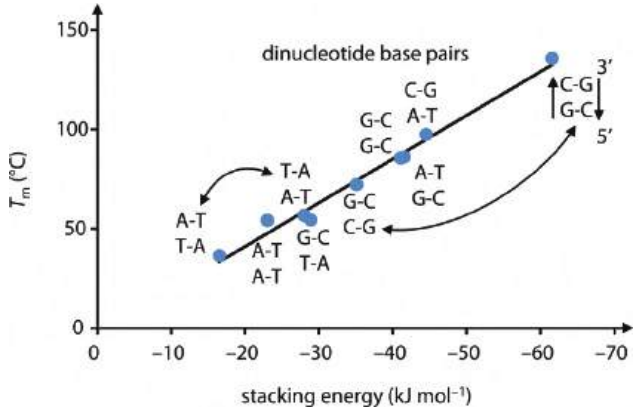


FIGURE 19.14: Linear correlation of calculated stacking energies with melting temperatures. T_m is calculated from the experimental melting temperature of oligonucleotides containing specific dinucleotides. The stacking energies were calculated using a potential function optimized for nucleic acids. A dash between bases marks a base pair that stacks onto another base pair. The arrows in the top right dinucleotide point from the 3'- to the 5'-end. T_m strongly depends on whether the purine is at the 5'- or 3'-end (curved arrows). Adapted from Ornstein & Rein (1978). Biopolymers 17(10): 2341–2360 and Gotoh & Tagashira (1981). Biopolymers 20(5): 1033–1042.

19.4.2 Base-pair geometries

A strictly planar base-pair stacking on top of another base pair should be the energetically most favorable situation because this way the largest surface is available for hydrophobic interactions. The optimal distance between base pairs is close to the sum of the van der Waals radii of the atoms making up the base pairs, about 340 pm. However, depending on the type of helical structure, the bases in a pair can be slightly twisted, sheared, opened, buckled, and staggered relative to each other (Figure 19.15). Some of these conformational adaptations are absolutely necessary to form the experimentally observed A-, B-, and Z-forms of DNA (see Section 19.5.1).

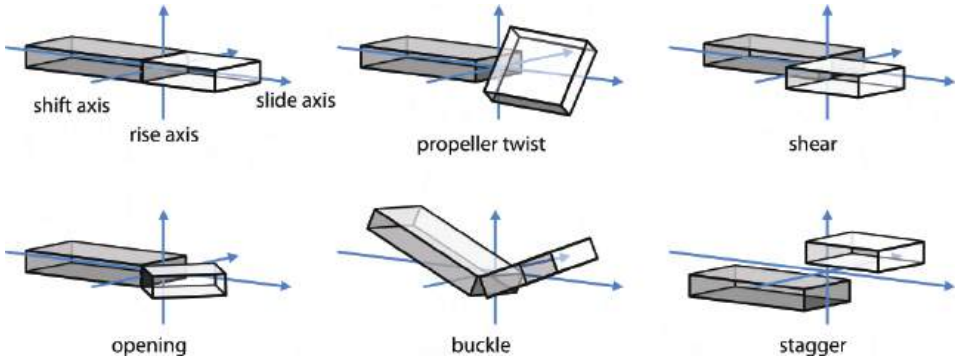


FIGURE 19.15: Base-pair geometries. The helix axis is the rise axis, the long axis in the base pair is the slide axis, and the shift axis is perpendicular to the first two. The movements are exaggerated for clarity. Bases can be laterally displaced, rotated, or tilted. A lateral displacement along the helix axis is the rise, displacements in the plane of the base pairs are shifts when they occur sideways with respect to the base pair or slide when they increase the distance between the bases of a pair. Shearing is a shifting of the bases, stagger is a change in rise. Rotations can occur around the helical axis (opening) or around the long axis of the base pair (propeller twist). A tilt out of the base plane (perpendicular to the helical axis) is a buckle.

The orientation of adjacent base pairs can also vary. These local structural adaptations of base-pair conformations, described by rise, slide, shift, twist, tilt, and roll (Figure 19.16), translate into fundamentally different structures of the various double helices.

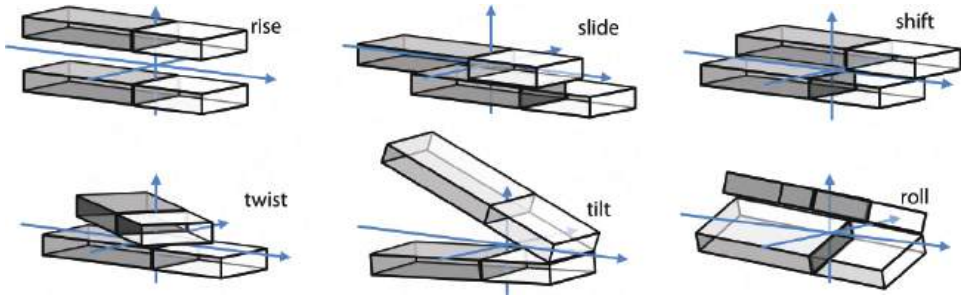


FIGURE 19.16: Relative orientations of base pairs. Rise, slide, and shift are displacements of one base pair relative to another along the same-named axes. Twist, tilt, and roll are rotations about the rise, shift, and slide axes, respectively.

19.5 DNA STRUCTURES AND CONFORMATIONS

DNA in its physiologically relevant forms adopts a number of fundamentally different, mostly helical structures. These include right- and left-handed double-stranded helices with their strands running anti-parallel. Although mostly rigid due to base pairing, unpaired regions in the middle (*breathing*) or at the ends (*fraying*) of a duplex may form spontaneously or can be induced by higher temperature or low ionic strength. DNA structures with three and four strands also exist, although their biological roles are not entirely clear. In addition to Watson-Crick base pairs, triplex and quadruplex DNA helices use the Hoogsteen edge of nucleobases to form their triple and quadruple base pairs (Section 19.5.2).

19.5.1 DNA DOUBLE HELICAL STRUCTURES

The archetype of the DNA structure is the double helix first described by Watson and Crick in 1953, with two complementary strands running antiparallel to each other. The strong stabilization by base-pair stacking causes even short duplexes to behave essentially as a single molecule. Double-stranded DNA is polymorphic and can adopt A-, B-, and Z-forms depending on the sequence and solvent conditions. A- and B-form helices are right-handed, but the Z-form is left-handed (Figure 19.17).

They also differ in their geometric parameters (Table 19.2). X-ray diffraction studies on isolated DNA fibers suspended in air established that the different forms of DNA depend on the humidity of the air. The canonical B-form of the DNA double helix is stabilized by high humidity, while higher salt concentrations, alcohols, and low humidity stabilize the other two forms. This means that under physiological conditions (in aqueous solution), the B-form prevails.

TABLE 19.2:
Helical parameters of double-strand nucleic acids.

structure	A	B	Z
helicity	right	right	left
helix diameter (nm)	2.3	2.0	1.8
mean # bp / turn, ⟨c⟩	11	10–10.5	12 (2 bp repeat)
twist / bp (°)	33	34–36	–9 and –51
displacement / bp (nm)	0.45	–0.02 to –0.18	–0.2 to –0.3
helix pitch (nm)	2.8	3.4	4.5
rise / bp (nm)	0.26	0.33–0.34	0.37
tilt of bp (°)	20	–6	–7
propeller twist (°)	18	16	0
glycosidic bond	*anti*	*anti*	*anti* (C) and *syn* (G)
sugar pucker	C3′-*endo*	C2′-*endo*	C2′-*endo* (*anti*) and C3′-*endo* (*syn*)
$P_n - P_{n+1}$ distance (nm)	0.56	0.66	0.56 and 0.64
C1′-C1′ distance (nm)	0.48	0.49	0.51 and 0.59
major / minor groove width (nm)	0.27 / 1.1	1.17 / 0.57	0.88 / 0.2
major / minor groove depth (nm)	1.35 / 0.28	0.88 / 0.75	0.37 / 1.38
average α and ζ (°)	–50 / –78	–41 / –157	–137 / 80 (Py) and 47 / –69 (Pu)
average χ (°)	–154	–102	–159 (Py) and 68 (Pu)

Py and Pu denote pyrimidines and purines, bp is short for base pair.

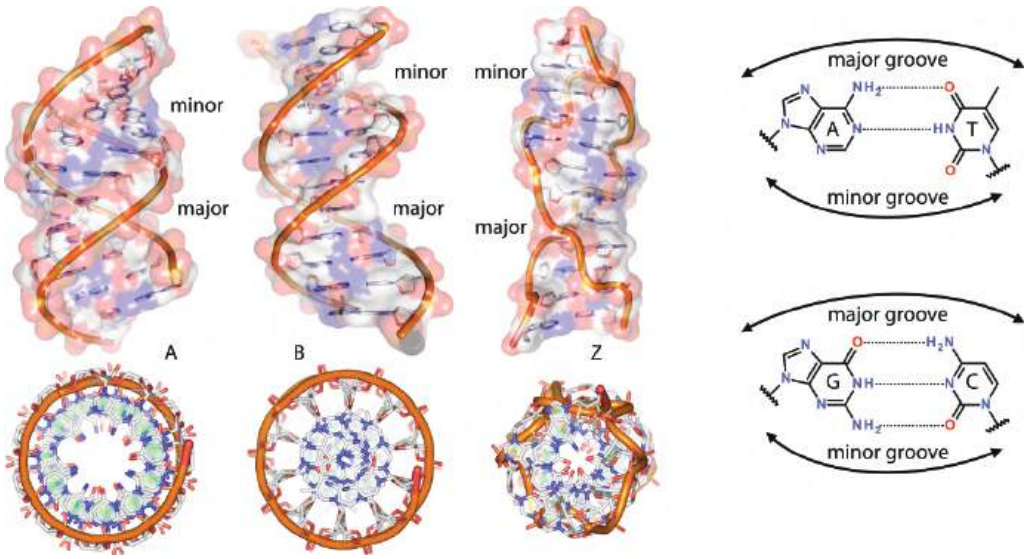


FIGURE 19.17: Double-stranded helices in A-, B-, and Z-form. A dodecamer sequence is shown for A-DNA, B-DNA, and Z-DNA (from left to right). A- and B-DNA are right-handed helices of 2.3 nm and 2.0 nm diameters, respectively. The narrow poly-CG Z-DNA (PDB-ID 4ocb) is left-handed. It earns its name from the zigzag path of the phosphate backbone. A Watson-Crick base pair from Z-DNA is shown in Figure 19.10. The major and minor grooves are indicated both on the overviews on the left and on the A-T base pair on the right. The upper (Hoogsteen edge) and the lower parts of the base pairs lie in the major and minor groove, respectively. Note that in A-DNA, the major groove is actually less wide than the minor groove. A-, B-, and Z-DNA are all antiparallel and are interconvertible.

All three main DNA conformations have been studied by fiber diffraction, NMR spectroscopy, and X-ray crystallography. The first atomic model of a DNA molecule was the B-DNA, built by James Watson and Francis Crick in 1953 using information from fiber diffraction photographs taken by Rosalind Franklin. B-DNA is the predominant conformation in solution, accounting for the vast majority of genomic sequences. The bases in B-DNA are not very much displaced from the helix axis (negligible slide) and almost perpendicular to it (tilt almost zero). Ten to 10.5 base pairs make up a full turn of the helix. The pitch is 3.4 nm, which corresponds to an optimal van der Waals distance (0.34 nm) between adjacent base pairs, enabling energetically favorable stacking interactions. All bases are in the *anti*-conformation, the deoxyribose is C2'-*endo*. The C3'-endo conformation would result in steric clashes and unfavorable interactions with neighboring base pairs. The surface of B-DNA shows two depressions known as *major and minor grooves*. The Hoogsteen edges of the nucleobases lie at the bottom of the wide and shallow major groove. Hydrogen bond donors and acceptors close to the glycosidic bond that are not involved in base pairing (the sugar edge) lie in the narrow minor groove (Figure 19.17) and provide a hydrogen-bond pattern for read-out by interaction partners (Section 19.5.4). X-ray crystallography and NMR have revealed a chain of water molecules bound in the minor groove of B-DNA, which is known as the *spine of hydration*. Both the major groove and the minor groove in B-DNA are sites of interactions with proteins and small molecules (Section 19.5.4).

A-DNA is the dominating conformation at low (75%) humidity and also the main conformation of double-stranded RNA. Similar to B-DNA, the base pairs in A-DNA and A-RNA are of Watson-Crick type, with all glycosidic χ-angles in the *anti*-conformation. In contrast to B-DNA, the ribose conformation is C3'-*endo*, not C2'-*endo*. In addition, the bases in the A-form are strongly displaced from the helical axis by about 0.45 nm, leaving an energetically unfavorable void along the center of the helix. This slide makes the minor groove very shallow and the major groove deep and narrow. Indeed, the minor groove in A-DNA and A-RNA is almost as broad as the major groove of B-DNA (Figure 19.17). The A-form has eleven base pairs in a single turn of 2.8 nm pitch. The vertical rise is thus only 0.26 nm per base pair, which is much smaller than van der Waals contact distance (0.34 nm). To avoid clashes, the bases in the base pairs are strongly tilted by ca. 20° relative to the helix axis.

Z-DNA has a repeating unit of two nucleotides, usually pyrimidine-purine, that regularly alternate in their glycosidic χ-angle between *anti* for the pyrimidine and *syn* for the purine. Along with the χ-angle, the sugar pucker alternates between C2'-*endo* (*anti*) and C3'-*endo* (*syn*), imposing a zigzag path on the phosphate backbone that gave the Z-form its name. Although twelve nucleotides make up a full turn, the Z-form is therefore better described as a repeating unit of six dinucleotides. Due to the smaller helix diameter, the ribose-phosphate backbone of the two strands is closer in Z-DNA than in A- or B-form DNA, leading to stronger electrostatic repulsion. When the electrostatic repulsion is minimized by counter ion binding, in the presence of high salt concentrations (4 M NaCl), or the ammonium ions of spermine and spermidine, the Z-DNA becomes stabilized. Certain chemically modified sequences containing 5-methylcytosine or 5-bromocytosine can form stable Z-DNA against which specific antibodies have been raised. Such antibodies have been shown to bind to the polytene chromosomes of enhanced transcription in *Drosophila* and near promoter sites of the c-*MYC* gene in mammalian nuclei, suggesting a biological role for Z-DNA. Indeed, stretches of Z-DNA may form to relax negatively supercoiled DNA (Section 19.5.3.2) during transcription of certain genes.

The formation of base-paired duplex structures is at the heart of biological and biophysical applications, ranging from gene silencing to surface attachment of molecules. For these applications, a high stability of the nucleic acids is desired, but both DNA and RNA are readily degraded by enzymes *in vivo*, and RNA is also prone to spontaneous hydrolysis. Stability against enzymatic or spontaneous hydrolysis can be increased by chemical modifications of the nucleic acid backbone. Examples of such modified nucleic acids are peptide nucleic acids (PNAs) and locked nucleic acids (LNAs) (Box 19.1). Complementary sequences of PNAs and LNAs can base-pair in the same way as natural nucleic acids (Section 19.4).

BOX 19.1: PEPTIDE NUCLEIC ACIDS (PNAs) AND LOCKED NUCLEIC ACIDS (LNAs).

PNAs and LNAs are backbone-modified derivatives of nucleic acids with improved nuclease and chemical stability that retain the base-pairing properties of natural nucleic acids. Complementary PNA and LNA strands form stable homo-duplexes and also hetero-duplexes with DNA and RNA. PNAs and LNAs are attractive candidates for sequence-specific DNA silencing and RNA interference.

PNAs are not acids but nucleobases connected to a neutral polyamide backbone (Figure 19.18). The whole ribose-phosphate backbone of DNA is replaced by a polymer of N-(2-aminoethyl)-glycine monomers or a similar molecule. There are two amide bonds per monomer. One is located close to the former position of the phosphodiester group, the other links a nucleobase to the polyamide backbone. Due to the absence of repulsive charges, PNAs can form exceptionally stable Watson-Crick base pairs and homo-duplexes. The lack of charges may also contribute to the high affinity of the hetero-duplexes formed with complementary RNA or DNA. The affinities are so high that even DNA and RNA sequences with abasic sites, i.e. sites lacking a nucleobase, are recognized. Almost irreversibly stable triplexes can form between single-stranded DNA and two PNA strands. Lacking phosphodiesters and having a non-polypeptidic backbone, PNAs are insensitive to nucleases and proteases. A drawback is that PNAs surmount membranes even less efficiently than natural nucleic acids. Liposomes (Section 18.4.3.2; Box 18.5) have been used to deliver them into cells.

LNAs are stable, nuclease-resistant RNA derivatives where the 2'-hydroxyl group is connected by a methylene (CH_2) group to the C4'-atom (Figure 19.18). The connection creates a rigid bicyclic ring system that locks the furanose into the C3'-*endo* conformation. In contrast to PNAs, LNAs retain a negatively charged ribose-phosphate backbone. LNAs bind to complementary RNA and DNA strands with very high affinity, presumably due to the rigidity of the modified ribose that reduces the entropic penalty for duplex formation. A drawback of the restricted flexibility of LNA compared to RNA is their limited structural variability (Section 19.6).

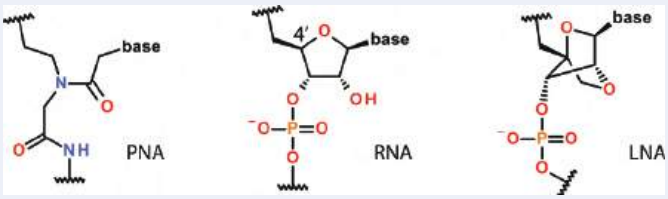


FIGURE 19.18: Comparison of peptide nucleic acids (PNAs) and locked nucleic acids (LNAs) with regular RNA. The RNA in the middle and the LNA on the right are shown in the 5'-3'direction. The neutral polyamide backbone of the PNA on the left has the same direction.

19.5.2 Triple and quadruple DNA helices

19.5.2.1 Triplexes

The major groove of B-DNA is wide enough to accommodate a third strand of DNA to form a triple helix or *triplex*. Especially when one strand of the duplex is rich in purines, a pyrimidine-rich third strand can read out the hydrogen-bonding pattern in the major groove generated by the Hoogsteen edges of the purines. Depending on the type of triplex, the nucleobases of the third strand form either Hoogsteen or reverse Hoogsteen interactions with the Watson-Crick base pairs of the duplex, leading to *base triplets* (Figure 19.19). Although they have been detected *in vivo*, it is uncertain

whether triple helices have a biological function. There are three types of DNA triple helices based on the sequence composition and the orientation of the third strand's backbone relative to that of the poly-purine strand of the duplex. In the *TC triplex*, a pyrimidine-rich third strand runs parallel to the purine-rich strand of the duplex, forming T·A-T and C^+·G-C Hoogsteen triplets (Figure 19.19). Triplets involving C^+·G-C pairings require low pH to protonate the N3-atom of cytosine (Figure 19.2). In a *GT triplex*, the third strand can be oriented either parallel or antiparallel to the poly-purine strand of the duplex, forming Hoogsteen or reverse Hoogsteen triplets, respectively, of the type G·G-C and T·A-T. In the *GA triplex*, the purine-rich third strand runs antiparallel to the purine-rich strand of the duplex, forming reverse Hoogsteen G·G-C and A·A-T triplets.

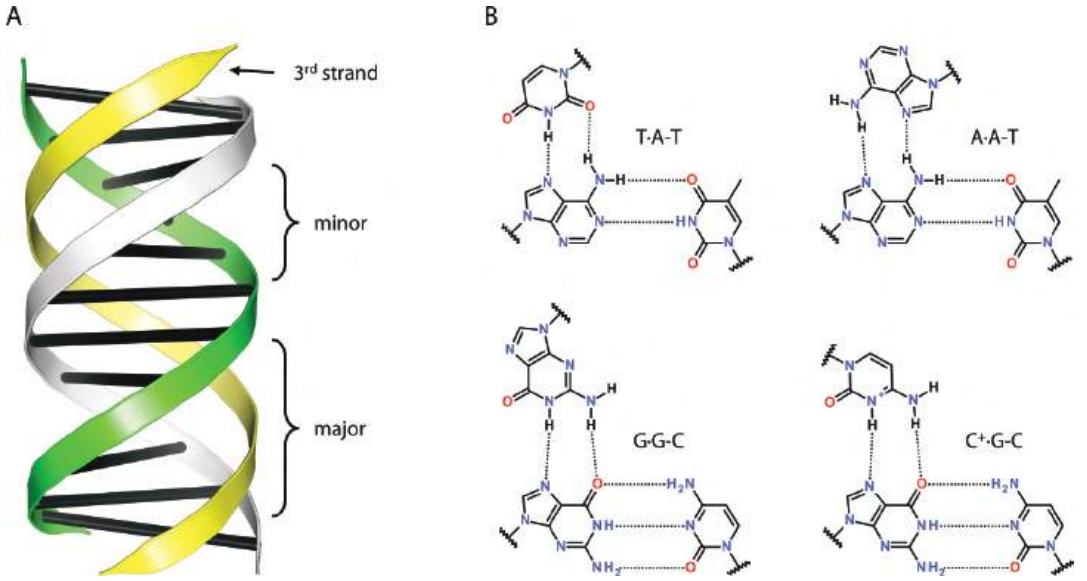


FIGURE 19.19: DNA triple helix. A: In the DNA triplex, an extra strand is bound in the major groove of B-DNA (colored yellow). B: Examples of base triplets observed in triple helices. Watson-Crick pairing is at the bottom, and (reverse) Hoogsteen pairings are at the top of the triplet. When cytosine forms a triple with a G-C pair (lower right), it must be protonated at the N3-atom. Only if there are two hydrogen bond donors can a Hoogsteen pair with the O6- and N7-atoms of guanine be formed.

Intermolecular triple helices have attracted much interest because of their therapeutic potential. For instance, oligonucleotides that are able to form triplexes with promotor regions in genomic DNA can be used for gene silencing. The triple helix approach can be used to target chemicals or proteins to specific sites on the DNA. To do so, crosslinking agents, nucleases, or transcription factors are fused to triplex-forming oligonucleotides. Once the triplex is formed, the cargo delivered by the oligonucleotide will crosslink, mutate, or hydrolyze the target DNA.

19.5.2.2 Quadruplexes and telomeres

DNA quadruplexes are formed by G-rich DNA strands, in which four guanines form base pairs using both their Watson-Crick and their Hoogsteen edges. The resulting planar *G-quartet* has almost completely saturated hydrogen bond donors and acceptors. The O6-atoms of the guanines point toward the center of the plane and can coordinate metal ions, which further stabilizes the quadruplex. The metal ions are sandwiched between two G-quartet planes and coordinated by eight O6-atoms from the guanines. Alkali ions such as K^+ and Na^+ are bound at the center of the quadruplex, but Sr^{2+} and Tl^+ ions can also bind *in vitro*. The direction of the strands in a quadruplex can be parallel or antiparallel with different connectivities when the quadruplex is made up of one, two, or four DNA molecules (Figure 19.20). In the nucleus, G-quadruplexes may form during transcription and replication when double-stranded DNA becomes transiently single stranded. G-quadruplexes may thus regulate gene expression and have a role in genome stability. G-quadruplexes also form in

telomeres, single-stranded G-rich DNA overhangs with repeat sequences (> 2000 times TTAGGG in vertebrates) at the ends of chromosomes where they may play a protective capping function. Outside the nucleus, G-quadruplexes can also form in mRNA and may be involved in translational control.

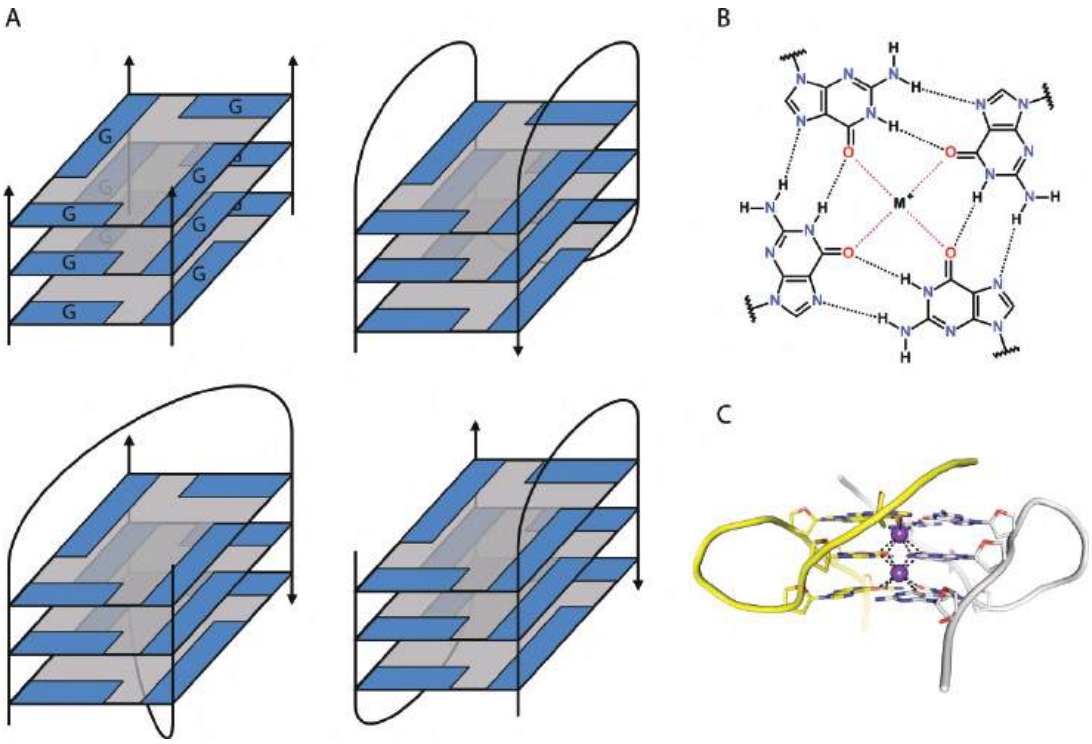


FIGURE 19.20: DNA quadruplexes. A: In the simplest case, four independent poly-G sequences form an intermolecular parallel quadruplex (top left). The quadruplex can also be formed by a single DNA strand (top right) or by two hairpin strands that either cross over (bottom left) or are parallel to each other (bottom right). B: Typical Hoogsteen G-quartet with a central metal ion M^+ (Na^+ or K^+). C: The side-view of a quadruplex made from two hairpins shows how K^+ is bound between neighboring quadruplets (PDB-ID 4p1d).

19.5.3 Higher-order DNA structures

DNA is often kinked, bent, extruded into *cruciform structures*, or curled up into *topoisomers*. The extent of deviation from the standard, linear B-DNA conformation is dependent on the actual sequence, the binding of protein partners, and the particular environment of the DNA. The formation of higher-order structures in aqueous solution is again driven by the minimization of hydrophobic surfaces and the maximization of favorable interactions. In these structures, the negatively charged backbones of different strands often come very close, and the resulting electrostatic repulsion needs to be minimized.

19.5.3.1 Helix junctions

Helix junctions are structures where several helical sections meet at a common point. Depending on the number of helices that meet at the junction, they are called four-way, three-way, and two-way junctions. The sequence at the junction determines the relative arrangement of the helical arms in helical junctions. Junctions often maximize stacking interactions by co-axial stacking of two or more helices. The global structure and conformational changes of helical junctions can be studied by FRET (Section 21.5.4).

Cruciforms and *Holliday junctions* are examples of four-way helical junctions. Cruciform structures can form in double-stranded DNA containing self-complementary *inverted repeat* or *palindromic* sequences. Under certain solution conditions, these sequences can rearrange to

extrude the DNA from its linear structure into a cross-shaped form resembling a four-way junction (Figure 19.21). The biological significance of cruciform structures is still debated because some of the nucleobases are unpaired, making the cruciform less stable than fully base-paired B-DNA. A biologically important structure that very much resembles a cruciform is the Holliday junction (Figure 19.21; Figure 19.22), named after Robin Holliday. In contrast to the cruciform structure, the Holliday junction is formed by four DNA strands, and lacks high-energy loop regions with unpaired nucleobases. The Holliday junction is an important intermediate in *homologous recombination*, a process used in cells to exchange genetic material. It is recognized by the RuvAB helicase, an ATP-dependent motor protein in bacteria that can move the junction along the DNA and determines the outcome of the recombination event.

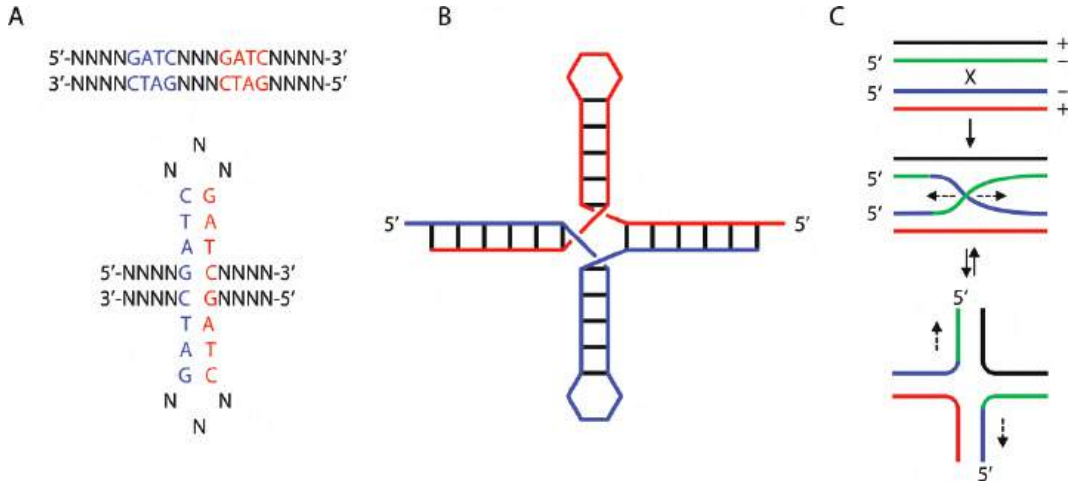


FIGURE 19.21: DNA cruciform structures. A: The short palindromic sequence GATC is separated by three nucleotides and flanked by four nucleotides. All nucleotides form base pairs in standard DNA, but here have been abbreviated as N for simplicity. Extrusion of the palindromic sequence into a cruciform leaves the bases between the inverted repeats unpaired. B: A four-way junction is formed by two DNA strands (red and blue). Base pairings are indicated by black lines. C: Formation of a Holliday junction during homologous recombination (marked by an X). The Holliday junction is made up of four DNA strands (different colors). After strand nicking and swapping of the nicked strand, the assembly can be redrawn to resemble a cruciform structure but without unpaired bases as in (A) and (B). The amount of DNA exchanged during the process depends on how far the junction moves along the sequence (indicated by dashed arrows).

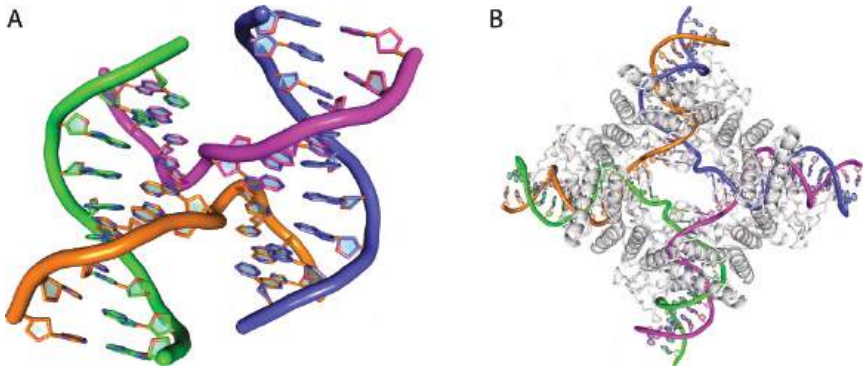


FIGURE 19.22: Crystal structures of Holliday junctions. A: The mini-junction (PDB-ID 467D) is made of the decamer sequence 5′-CCGGGACCGG-3′, which is self-complementary but for the central GA sequence. These bases form non-standard base pairs at the center of the junction. The four helical arms of the junction form two duplexes by pairwise co-axial stacking. B: By contrast, the Holliday junction of the LoxP sequence 5′-ATAACTTCGTATA-NNNTANNN-TATACGAAGTTAT-3′ is a planar cruciform-like structure stabilized by the Cre protein.

The replication fork is an example of a three-way junction: the DNA that has not yet been replicated forms one helical arm, the two copies of newly synthesized DNA form the other two arms. If DNA replication stalls due to a DNA lesion, there are two mechanisms for repair: homologous recombination or the transient formation of a four-way junction termed *chicken foot structure*. In this structure, the newly synthesized strands can serve as their own templates until enough DNA is synthesized to bypass the original lesion. Reversal of the chicken foot structure to a three-way junction restores the replication fork (Figure 19.23).

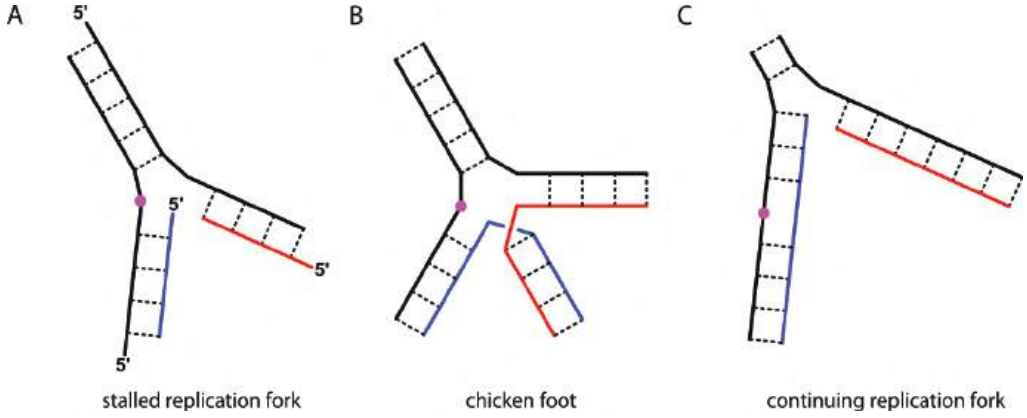


FIGURE 19.23: Junctions in DNA replication. The parent strands are shown in black with base pairings as dashed lines. A DNA lesion is marked by a magenta circle. A: Intact replication fork. The 5′-ends of the parent DNA, the continuously synthesized strand (red), and the Okazaki fragment (blue) are indicated. Replication stops at the lesion. B: The newly synthesized strands use each other as new template by transiently forming a structure that resembles a chicken foot. C: The chicken foot structure is reverted into a replication fork, and the DNA lesion has been bypassed.

Helical junctions are also frequent structural elements in RNA, and all aspects discussed here for DNA junctions also pertain to RNA junctions. A prominent example of an RNA four-way junction is tRNA (Section 19.6.2).

19.5.3.2 DNA supercoiling

In the schematic depiction of the DNA strands in a Holliday junction (Figure 19.21) and the replication fork (Figure 19.23), we have neglected the helical structure of DNA. The double-stranded helical structure of DNA is an impediment to essentially all biological processes that require strand separation. When DNA is replicated, the two strands have to be separated to form the replication bubble. The movement of the replicating DNA polymerase along the DNA and the concomitant separation of the two strands of the double helix require the two DNA strands to rotate. However, this rotation is not feasible in cellular DNA that is packed very tightly. In circular covalently closed DNA such as plasmids, bacterial chromosomes, and mitochondrial DNA, the ring shape fixes the ends of the DNA and prevents their rotation. In eukaryotes, DNA is bound to nucleosomes (see Figure 19.28) that form higher-order structures. Chromosomal DNA can therefore be regarded as an assembly of circular domains, loops whose ends are fixed and cannot rotate freely. The same is true for very long linear DNA, which is too inert to freely rotate about its long axis.

As the DNA cannot freely rotate, movement of the replication fork along the DNA double helix generates torsional stress. This torsional stress leads to overwinding of the DNA ahead of the replication fork and underwinding behind it. The same is observed when the RNA polymerase moves along the DNA in the transcription bubble. Overwinding and underwinding of the double helix lead to coiling of the helix around itself, and to the formation of supercoils. DNA that is overwound forms positive supercoils that stabilize DNA in its double-stranded form and counteract strand separation. Overwinding ahead of the replication fork or the transcription bubble thus impedes fork progression and replication, and needs to be removed. Underwound DNA forms negative supercoils.

Underwinding favors strand separation, which is the reason for the higher propensity of negatively supercoiled DNA to form cruciform structures *in vitro*.

The helical nature of the DNA also has consequences at the end of replication: the two DNA circles generated by replication of a circular DNA are linked like two members of a chain (Figure 19.24). These linked circles are called *catenanes*. For segregation of one copy of the DNA to each daughter cell during cell division, catenanes have to be resolved. Other knotted forms of DNA are also observed, but they will not be treated here.

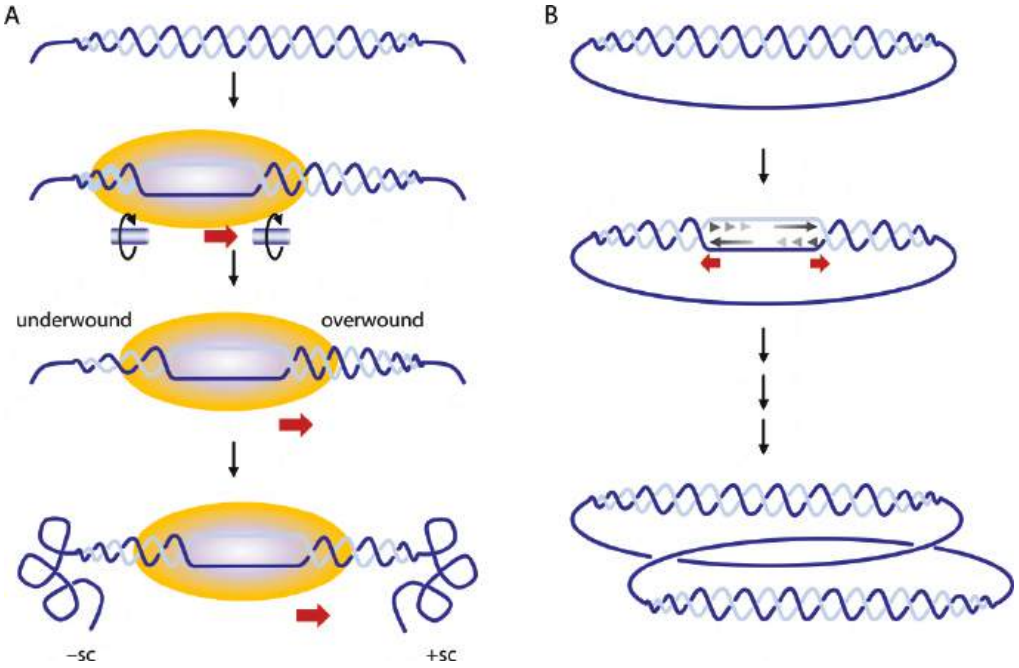


FIGURE 19.24: Movement of RNA polymerase and transcription bubbles generates supercoiling. A: DNA transcription by RNA polymerase (orange) requires unwinding of the DNA double helix and generation of a transcription bubble. During transcription, movement of the transcription bubble requires unwinding of the DNA in the direction of RNA polymerase movement and rewinding behind it. Because the ends of a circular DNA cannot rotate freely, the torsional stress leads to positive supercoiling of the DNA ahead of the RNA polymerase, and negative supercoiling in its wake. The positive supercoils ahead of the moving RNA polymerase are an impediment for its progression, and need to be removed by topoisomerases. B: DNA replication of covalently linked DNA circles leads to the production of catenanes, two linked double-stranded DNA rings. The two strands of the template DNA are depicted in blue, the newly synthesized DNA in gray. The arrow represents leading strand synthesis, and the triangles indicate Okazaki fragments in lagging strand synthesis. For segregation of the two DNA molecules to the daughter cells, the catenanes are unlinked by topoisomerases.

To describe supercoiling quantitatively, we first define the *linking number* (*Lk*) of a DNA molecule. We have seen in Section 19.5.1 that the average number of base pairs per turn in B-DNA is $\langle c \rangle$ = 10–10.5. The exact number depends on temperature, ionic strength, and pH. B-DNA with 63 base pairs thus forms six helical turns. We can directly link the ends of this DNA into a covalent circle. The linking number of the circular DNA describes how often the strands are linked within this circle. In this case, the linking number is simply the number of base pairs divided by the number of base pairs per turn, *i.e.* the number of helical turns (*Lk* = 6). Such a DNA is called *relaxed* DNA, and its linking number is also referred to as Lk^0. The linking number is always an integer number.

We now take the same piece of DNA with six helical turns, keep one end fixed and rotate the other end in the rotational sense of the helix, such that one helical turn is removed (Figure 19.25). The number of helical turns then reduces to 5, and the number of base pairs per turn increases: the DNA is said to be *underwound*. Conversely, we can *overwind* this DNA by rotating one end against

the rotation sense of the helix, such that the number of helical turns increases. Overwinding and underwinding DNA cause torsional strain in the DNA. When the overwound or underwound DNA is covalently closed, this strain can be accommodated in two ways: either by changes in twist *Tw* or by changes in writhe *Wr*. The twist *Tw* corresponds to the number of turns of the double helix, the writhe *Wr* quantifies how often the double helix is coiled (Figure 19.25). In other words, the writhe quantifies the number of supercoils present. *Wr* can be positive or negative, reflecting the number of positive or negative supercoils present in the DNA. The double-stranded crossovers in positively supercoiled DNA are positive, and the resulting superhelical turn is left-handed. Negatively supercoiled DNA has negative crossovers and right-handed superhelical turns (Figure 19.28).

The twisting number *Tw* and the writhe *Wr* add up to the linking number *Lk*:

$$Lk = Tw + Wr \qquad \text{eq. 19.3}$$

While *Lk* is always an integer number, *Tw* and *Wr* can adopt any value. Relaxed B-DNA is described by $Wr = 0$ and $Tw = Lk = N/10.5$, where *N* is the number of base pairs. Chemically identical DNA molecules that differ only by their *Lk* are called *topoisomers* and are interconverted by enzymes termed *topoisomerases*.

So far we have restricted our considerations to DNA molecules with integer numbers of helical turns. What happens when the number of base pairs *N* is not a multiple of 10.5 base pairs per turn, and the ratio $N/\langle c \rangle$ gives a non-integer value? In this case, covalent joining of the ends into a circle is not possible by mere bending of the DNA, but requires some torsion of the DNA to align the ends: we have to slightly over- or underwind the DNA double helix for ring closure. The linking number of the relaxed DNA circle will be the closest integer to the calculated value ratio $N/\langle c \rangle$. For a 108 base-pair DNA, $N/\langle c \rangle = 10.3$. This DNA will form a covalently closed circle with $Lk = 10$. The deviation of the linking number *Lk* from $Lk^0 = N/\langle c \rangle$ gives rise to some torsional stress even in the relaxed form. As a consequence, the most relaxed topoisomer of this DNA already has some twist and/or writhe. While this torsional stress is small for large DNA molecules, it becomes more and more pronounced for small DNA circles.

The linking number of covalently closed DNA is constant. *Lk* can only be changed by transient breakage of the phosphodiester backbone. A change in linking number generates a new topoisomer that will adjust to its different linking number either by over- or underwinding (twist), or by supercoiling *Wr*, or both. Starting from the relaxed DNA circle with $Wr = 0$ and $Tw = Lk = 6$ (see above), a negative supercoil can be introduced by cleavage of both strands, introducing underwinding ($Tw = 5$ turns) and re-ligation of both strands ($Lk = 5$). The twist *Tw* either remains at 5, or increases to 6 upon formation of one negative supercoil ($Wr = -1$). Conversely, the relaxed DNA can be positively supercoiled by cleavage, overwinding, re-ligation, and formation of one positive supercoil ($Lk = 7$, $Tw = 6$, $Wr = +1$).

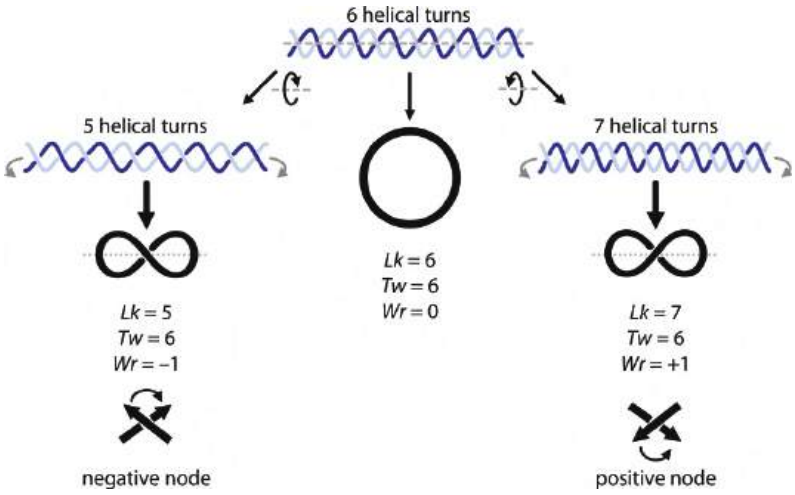


FIGURE 19.25: DNA supercoiling. A regular B-form DNA with 6 helical turns can be circularized to a covalently closed, relaxed DNA (middle, circle). The linking number of the circular DNA is $Lk = 6$. Underwinding (left) or overwinding (right) before circularization introduces torsional stress that is distributed into twist *Tw* and writhe *Wr*, and generates negatively and positively supercoiled DNA. The dotted gray line marks the double-helical and superhelical axis, respectively. Negative supercoils form negative crossovers (alignment of the upper double helix with the bottom double helix requires a rotation in clockwise direction). Positive supercoils form positive crossovers (alignment of the upper double helix with the bottom double helix requires a rotation in counter-clockwise direction).

Topoisomers have different hydrodynamic radii and can therefore be resolved by gel electrophoresis (Figure 19.26). Negative and positive topoisomers of the same linking number have different propensities to interact with intercalators, which can be used for their separation by gel electrophoresis in two dimensions. Intercalators are planar, aromatic molecules that bind to DNA by insertion between two adjacent base pairs (Section 19.5.4.2). A well-known intercalator is ethidium bromide (see Figure 19.33) that becomes fluorescent upon DNA binding and is used to visualize DNA in gels. Intercalation leads to local unwinding of the DNA and a decrease in *Tw*. For circular DNA, *Lk* is constant, and a decrease in *Tw* inevitably leads to an increase in *Wr* (eq. 19.3). A negatively supercoiled plasmid thus becomes more and more relaxed when intercalators bind. Once it is fully relaxed, further binding of the intercalator leads to accumulation of positive supercoils until the maximum number of supercoils is reached, and the DNA is saturated with intercalating molecules. The more negatively supercoiled the plasmid is, the more intercalating molecules it can bind. A positively supercoiled plasmid also becomes more and more positively supercoiled with each molecule that intercalates, and reaches the limiting number of supercoils with fewer intercalating molecules bound. As a consequence, negatively supercoiled DNA has a lower electrophoretic mobility than positively supercoiled DNA in the presence of intercalators. If the second dimension of gel electrophoresis is performed in the presence of intercalators such as chloroquine, negatively supercoiled DNA lags behind, and can be separated from positively supercoiled DNA with the same *Wr*. A ladder of different topoisomers is separated into an arch with a left branch corresponding to the negatively supercoiled species and a right branch formed by the positively supercoiled DNA.

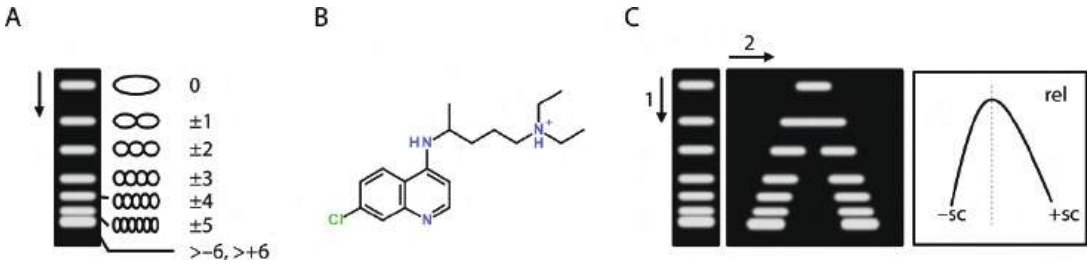


FIGURE 19.26: Separation of DNA topoisomers. A: The higher the number of supercoils the more compact the molecule, and the faster it migrates on an agarose gel. B: Chloroquine is a DNA intercalator that causes untwisting of the DNA upon binding. C: Negatively and positively supercoiled DNA can be separated by two-dimensional gel electrophoresis. After separation in the first dimension according to the hydrodynamic radius and thus to the number of supercoils, the separation in the second dimension is performed in the presence of an intercalator such as chloroquine. The more negatively supercoiled the DNA, the more intercalating molecules are bound, leading to a decrease in electrophoretic mobility in the second dimension. The more positively supercoiled the DNA, the fewer intercalating molecules are bound, and the electrophoretic mobility remains high. Separation of DNA with the same number of negative and positive supercoils in the second dimension leads to an overall arch-like band pattern. The left branch corresponds to negatively supercoiled DNA, the right branch to positively supercoiled DNA.

Supercoiled DNA has a higher free energy than relaxed DNA because of its torsional energy (Figure 19.27). DNA supercoiling is thus an endergonic reaction. Relaxation, on the other hand, is an exergonic process and thus in principle a spontaneous reaction. Nevertheless, the thermodynamically favorable relaxation of DNA by rotating about its ends is not possible if the DNA is long and too inert to rotate, or if the DNA is circular covalently closed. In the cell, the interconversion of topoisomers is catalyzed by topoisomerases that transiently cleave the DNA to change its linking number (Figure 19.27).

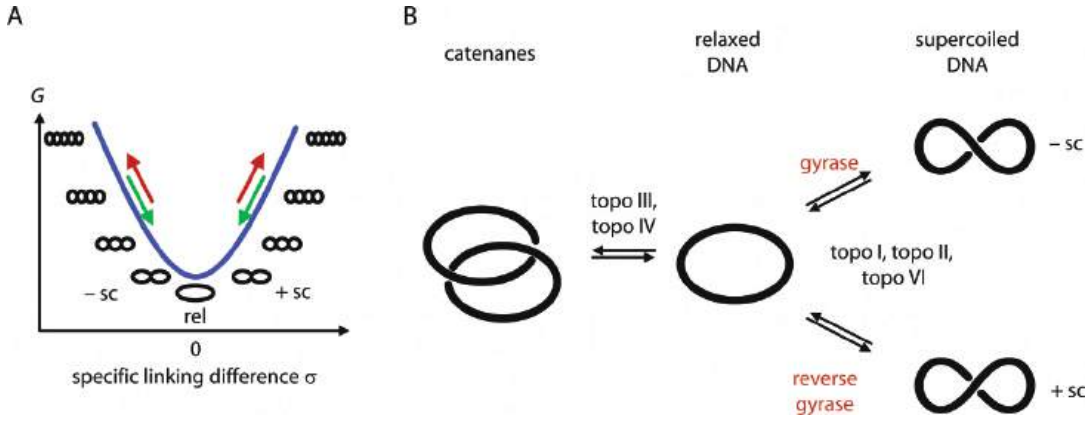


FIGURE 19.27: Interconversion of topoisomers by DNA topoisomerases. A: Energy diagram. The torsional energy of supercoiled DNA and the free energy *G* increase with the number of supercoils. Relaxation of DNA is an energetically favorable process (green), whereas the introduction of supercoils is endergonic (red). B: DNA topoisomerases interconvert topoisomers. Topo I, III, and reverse gyrase are type I topoisomerases that cleave one strand of their double-stranded DNA substrate. Topo II, IV, VI, and gyrase are type II topoisomerases that cleave both strands of the DNA. Topo II and IV catalyze DNA catenation and decatenation. Gyrase and reverse gyrase introduce negative and positive supercoils into DNA, respectively, at the expense of ATP hydrolysis (red). Topo I, II, and IV catalyze DNA relaxation. Supercoiled DNA can also be catenated/decatenated (not shown).

In vivo, topoisomerases constantly introduce and remove supercoils. The steady-state supercoiling level of cellular DNA is determined by the opposing activities of the set of topoisomerases present in the organism. The overall degree of supercoiling is quantified by the *specific linking difference* σ of the DNA.

$$\sigma = \frac{Lk - Lk^0}{Lk^0} = \frac{\Delta Lk}{Lk^0} \qquad \text{eq. 19.4}$$

Lk^0 is the linking number of the relaxed state of the DNA (N/10.5), Lk is the actual linking number, and ΔLk the linking difference. The parameter σ is also called the *superhelical density*, although it does not specify the distribution of twist and writhe. The *specific linking difference* is normalized to the number of helical turns (Lk^0) and thus the size of the DNA, and allows comparing the superhelicity of different DNAs. Positive values of σ correspond to positively supercoiled/overwound DNA, negative values to negatively supercoiled, underwound DNA. Using eq. 19.4, we can calculate the effect of deviations between Lk and Lk^0 for DNAs whose number of base pairs is not an integer multiple of 10.5: for a 108 bp DNA, Lk^0 is 108/10.5 = 10.3. The covalent circle has a linking number Lk = 10, which gives ΔLk = −0.3, and σ = −0.3/10.3 = −0.03. The same linking difference for a 3 kb plasmid corresponds to σ = −0.001, and the torsional stress of the most relaxed species is negligible. Most organisms maintain their DNA moderately negatively supercoiled. The local supercoiling density is a major determinant for gene expression.

19.5.3.3 DNA bending and kinking

Both bending and kinking change the direction of the axis of the DNA double helix. DNA kinking is brought about by an abrupt change in local backbone conformation at a specific base pair. In contrast, bending is defined as a continuous change in curvature over a stretch of several base pairs. Both bending and kinking reduce the stacking between adjacent bases, which is energetically unfavorable. Kinks and bends are often induced and stabilized by protein binding.

Short sequences of 4–6 A-T base pairs (*A-tracts*) can induce a curvature of ca. 18°±2° in B-DNA. The curvature is caused by a decreasing width of the minor groove from the 5'- to the 3'-direction along the A sequence. This decrease in turn is a result of the large negative propeller twist and the increasing negative inclination of the A-T base pairs, giving rise to bifurcated hydrogen bonds between the N6-atom of an adenine and the O4-atom of a thymine in the next A-T base pair in the 3'-direction. If the A-tracts repeat in phase with the rise of the B-DNA (10–11 base pairs), they are all on the same side of the helix and the curvatures induced by each A-tract sum to a global curving of the DNA. Curvature can be detected by methods that are sensitive to the hydrodynamic radius, including light scattering (Section 23.1), analytical ultracentrifugation (Section 29.2), gel permeation chromatography, and electrophoresis. Poly-A tracts are often found in regulatory elements of DNA, suggesting a role in gene transcriptional regulation, replication, or recombination. *In vitro*, intrinsic bending of DNA by phased A-tracts can be used to create small DNA rings of about 100 base pairs.

Functional bending and kinking of DNA can also be induced or strengthened by protein binding. An example is the cellular storage form of DNA, *chromatin*, where DNA is wrapped around a core particle of eight histones to form a *nucleosome*. The histone octamer wraps a stretch of 146 base pairs of regular B-DNA on its outer surface in almost two complete turns of a left-handed superhelix (Figure 19.28). The driving force for bending is the electrostatic attraction between basic residues on the surface of histones and the negatively charged ribose-phosphate backbone of DNA. In fact, the electrostatic surface potential of the histone octamer has a helical pitch that is followed by the phosphate backbone. Bending of the DNA around the nucleosome compresses the minor groove, leading to an increase of negative charge density (*electrostatic focusing*) that is neutralized by arginine side chains on the surface of the histones. The C-terminal domain of gyrase (Section 19.5.3.2) has a similar electrostatic helical tract to bend and wrap DNA.

An example of protein-induced DNA kinking is the DNA bound to the TATA-box binding protein (TBP). An estimated 10–20% of eukaryotic genes have a TATA sequence upstream of their promoter. TBP is a transcription factor that assists in positioning the RNA polymerase II for initiation of transcription. Binding of TBP to the TATA box kinks the DNA twice, once at each

TA sequence, by a total of about 115° (Figure 19.28). DNA unwinding and kinking is facilitated by the comparatively weak stacking interaction of the 5'-TA-3' sequence (Figure 19.14). A pair of Phe residues, one on each side of the TATA box, wedges between the base pairs of the 5'-TA-3' sequence. One of the Phe side chains stacks on the adenine, the other packs on the ribose of the complementary T nucleotide. Stacking of aromatic protein side chains on exposed nucleobases is a general principle of protein-DNA interaction and often seen in complexes where the DNA is kinked. The new stacking interaction formed compensates for the loss of stacking between the two bases flanking the kink. In the TBP-TATA box complex, the large concave 10-stranded β-sheet of TBP covers and stabilizes the widened minor groove using mainly hydrophobic interactions between protein side chains and the ribose furan rings. The extreme widening of the minor groove upon TBP binding changes the overall conformation of the B-DNA to A-DNA. We will encounter more examples of DNA bending and kinking upon protein binding in the next Section (19.5.4.1; Figure 19.31) where we discuss the structural principles of DNA-protein recognition.

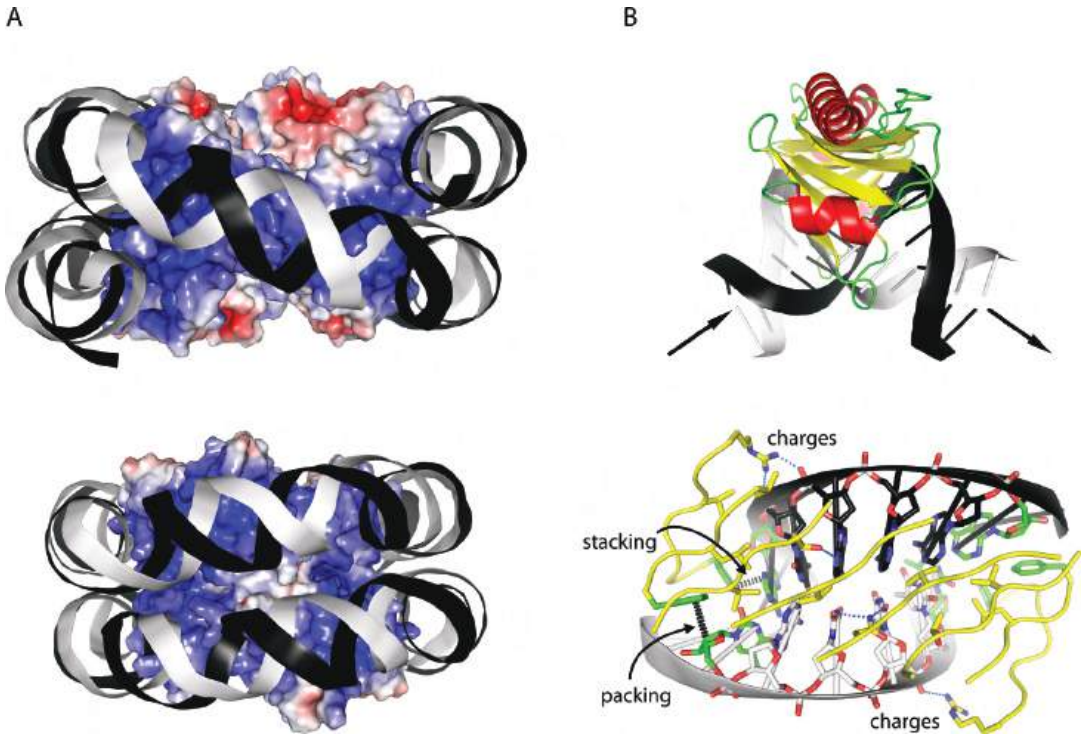


FIGURE 19.28: Protein-induced DNA bending. A: In the nucleosome, the histone octamer wraps a 146 base pair double-stranded DNA (PDB-ID 1kx5). The octamer is shown as a surface electrostatic potential with positive potential colored in blue. The trace of the positive potential on the octamer drives electrostatic bending of the double-stranded DNA (black and grey strands). The top view highlights the left-handed DNA superhelix. The bottom view is a 180° rotation about the vertical axis showing two traces of the positive potential that are followed by the double-stranded DNA. B: Strong kinking of DNA by the TATA-box binding protein (PDB-ID 1cdw). Arrows in the top panel give the direction of the DNA helix axis before and after kinking. The bottom panel is a view through the protein to the widened minor groove where van der Waals interactions (yellow side-chains) dominate. The two pairs of Phe residues (green) responsible for the kink pack on the adenine and a ribose. It can be seen that the affected TA base pair (green) is not planar any more.

19.5.4 DNA INTERACTIONS WITH PROTEINS AND LIGANDS

The driving forces for protein-DNA and ligand-DNA interactions are quite the same as for protein-protein and protein-ligand interactions, namely maximization of non-covalent interactions and minimization of exposed hydrophobic surfaces. Interactions include hydrogen bonds for specificity,

cation binding to neutralize the negative charges of the phosphate backbone, and van der Waals contacts between hydrophobic surfaces.

19.5.4.1 DNA recognition by proteins

Binding of proteins such as transcription factors and other regulators of gene expression to DNA can be sequence-specific or unspecific. Sequence-specific binding usually involves direct hydrogen bonds between the nucleobases and protein side chains, a recurring principle of molecular recognition. Sequence-independent or indirect DNA recognition, on the other hand, does not rely on direct contact with base pairs but on variations in the electrostatic potential and the shape of the DNA. An example for sequence-independent DNA binding is the nucleosome that stores genomic DNA by wrapping the negatively charged double-strands around a positively charged core of histone proteins (Figure 19.28). Other proteins that bind DNA (and RNA) independent of its sequence often have intrinsically disordered regions with positively charged side chains. These extended regions may act as an electrostatic "lasso" that helps to capture DNA. Once in complex with DNA, the extended region can fold into a defined structure for increased affinity. Finally, Z-DNA can be recognized in a sequence-independent manner *via* its peculiar zigzag electrostatic potential: a domain in the ADAR1 protein (double-stranded RNA adenosine deaminase) recognizes the charge and shape of five consecutive phosphodiester groups along the Z-DNA backbone. Most biological protein-DNA interactions are, however, sequence-specific.

Single-stranded nucleic acids are recognized sequence-specifically by hydrogen bonds of protein main-chain and side-chain groups to the Watson-Crick edges of nucleobases. Such single-stranded regions may form upon bending, kinking, or partial unwinding of double helices. Sequence-specific recognition of intact double helices is possible through hydrogen bonds to groups in the major groove, outside the Watson-Crick edges. The side chains of proteins that bind to the major groove form hydrogen bonds with the Hoogsteen edges of the nucleobases. Regardless of their architecture, sequence-specific DNA binding proteins use common themes to read out the nucleotide sequence (Figure 19.29).

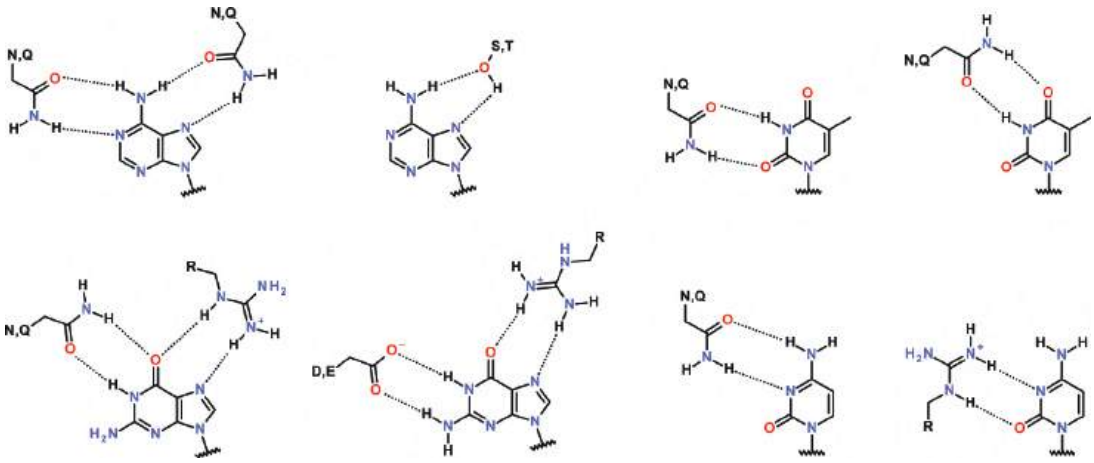


FIGURE 19.29: Hydrogen bonds between nucleobases and protein side chains. Amino acids are given in single letter notation. Both the Watson-Crick and the Hoogsteen edges are used for hydrogen bonding and recognition. For arginine, several orientations of the guanidinium side chain relative to the nucleobase are possible.

Proteins recognize DNA sequences either through the minor or through the major groove. The most prevalent motif of sequence-specific DNA recognition is an α-helix binding to the major groove. The side chains on one side of the α-helix recognize the DNA sequence by binding to the Hoogsteen edge of a stretch of nucleobases in the major groove. This binding mode is found in *helix-turn-helix*, *basic leucine zipper*, and the *zinc finger* proteins (Figure 19.30). A minimal helix-turn-helix or HTH motif has about 20 residues and folds into two α-helices connected by a β-turn.

Variations of the HTH motif with a longer loop are termed *helix-loop-helix motif*. In both motifs the two helices are almost perpendicular to each other. The second helix, termed the *recognition helix*, binds into the major groove of DNA in a sequence-specific manner. The HTH motif on its own is not stable, but is always part of a larger protein that may contribute additional residues for DNA binding.

The basic leucine zipper (bZIP) motif consists of two amphipathic α-helices forming a left-handed coiled-coil super-secondary structure (see Figure 18.24) at one end and extend uninterrupted into the major groove of DNA at the other end. bZIP proteins are therefore examples where a single α-helix has two different functions, dimerization and DNA binding (Figure 19.30). Transcription factors with this motif can be homo- or heterodimeric. Interestingly, in the Jun/Fos heterodimeric transcription factor, the monomers are unfolded but form a Leu-zipper in the heterodimer. The basic DNA-binding sequences in transcription factors, too, are unfolded unless they bind to DNA.

Zinc fingers (ZIFs) are small domains of up to 90 but generally less than 50 residues that are stabilized by a Zn^{2+} ion in a tetrahedral coordination. Zinc fingers may be classified based on their Cys/His composition, which includes Cys2His2 (C2H2), Cys2HisCys, and Cys4. Among these, the C2H2-type zinc fingers are the archetype of the family and the most studied. C2H2 zinc fingers are present in hundreds of transcription factors and coordinate Zn^{2+} between a two-stranded β-sheet and an α-helix using two Cys and two His side chains. The α-helix of the C2H2 zinc finger binds to the major groove. Up to sixty copies of this type of zinc finger are found in transcription factors. Because each helix recognizes 3–4 adjacent base pairs, each additional zinc finger increases DNA specificity. Artificial concatenation of zinc fingers with different sequence specificities allows generation of DNA-binding modules capable of recognizing a single site in the whole genome, which is used for gene editing (Box 19.2).

DNA binding by HTH, bZIP, and ZIF proteins generally does not disrupt base pairing. Instead, the DNA is usually bent to different extents (Figure 19.30).

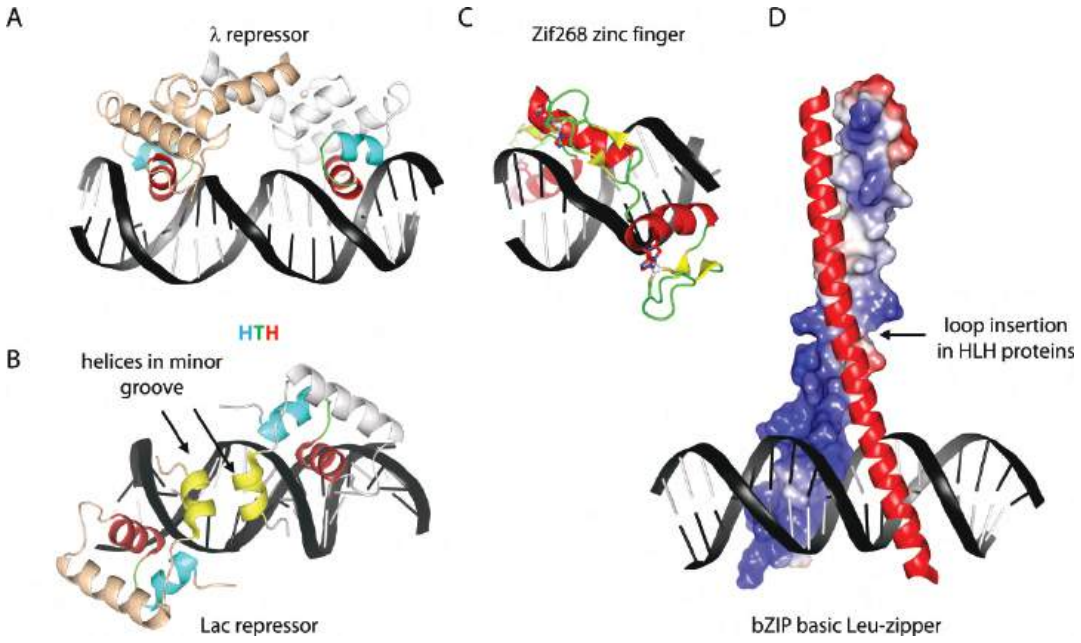


FIGURE 19.30: α-helical DNA binding motifs. A: The λ-repressor (PDB-ID 1lmb) is a dimeric protein that binds to the operator of the *cro* gene of the viral DNA to maintain a lysogenic (silent, non-virulent) life cycle. The recognition helices (red) are 3.4 nm apart, corresponding to a full turn of B-DNA. B: The Lac repressor, which shuts down transcription of the *lac* operon when bound to its operator DNA (PDB-ID 1efa), not only binds to the major groove with an HTH motif, but also strongly distorts the minor groove of the DNA by additional α-helices (yellow). C: In the eukaryotic transcription factor Zif268, three consecutive zinc fingers, spaced at intervals of 3 base pairs, bind the major groove with α-helices (PDB-ID 1aay). D: bZIP proteins also recognize DNA by way of an α-helix in the major groove. One of the amphipathic helices of the dimeric viral Jun protein is shown with its electrostatic surface potential to highlight the basic DNA-binding region and the hydrophobic leucine zipper (PDB-ID 2h7h).

In contrast to the major groove, the diversity of hydrogen bond donors and acceptors in the minor groove is not high enough to enable sequence-specific binding of proteins. In rare cases, the minor groove can also host an α-helix from an HTH motif or another DNA-binding motif. The Lac repressor is such an example where the minor groove is distorted so much that two α-helices (not belonging to an HTH motif) fit into it side-by-side (Figure 19.30). The minor groove of DNA either interacts with a single β-strand or loop or, after widening and DNA bending or kinking, with a whole β-sheet as in the case of TBP (Figure 19.28).

BOX 19.2: GENE EDITING BY PROGRAMMABLE ZINC FINGER NUCLEASES.

Many human diseases are the result of point mutations in genes. Traditional gene therapy seeks to repair the lost function of the mutated gene by adding another copy, often using a virus as a vector. The best way to treat a genetic disease would be to restore the wild-type sequence in the affected gene in the germline. This requires precise cleavage of the DNA double-strand of the human genome ($3.2{\cdot}10^9$ base pairs) at the location of the mutation, followed by elimination of the mutation by cellular repair mechanisms. High cleavage specificity can be achieved with *programmable zinc finger nucleases* (ZFN), chimeric enzymes with a specific DNA-recognition and a DNA-cleaving part. DNA recognition is achieved through a sequence of several zinc fingers. Each zinc finger recognizes three base pairs. Zinc fingers of different specificities can be concatenated in a modular fashion to recognize virtually any DNA sequence, which is important for specifically and exclusively cleaving the genome at the target site. A series of four zinc fingers is sufficient to address a single site in the human genome. The zinc fingers are fused to the DNA-cleavage domain from the bacterial restriction endonuclease FokI that lacks sequence specificity. The zinc fingers anchor FokI at a specific sequence on the DNA; cleavage by FokI then generates a double-strand break with overhangs. If an oligonucleotide with the healthy gene sequence is provided together with the ZFN, homologous recombination by the cell can replace the mutated region with the correct one. The technique has been used for gene editing (mutation, deletion, insertion) in animals and plants. The CRISPR-Cas9 system uses a similar principle. Here, the specific DNA-recognition part is an RNA complementary to the target site, the DNA-cleaving part is the bacterial nuclease Cas9 (Box 19.4).

We have seen before that major-groove binders recognize the DNA through α-helices. Homeobox proteins are an example of a hybrid DNA-binding protein. They contain a canonical HTH motif that interacts with the major groove, but in addition they bind into the minor groove with a loop region extending away from the protein.

The major groove can also host small β-sheets. The MetJ repressor is the prototypical representative of the β-hairpin/ribbon class of DNA-binding proteins. Its six families of divergent structures all use β-sheets for DNA recognition. MetJ is a homodimer with a protein-protein interface constructed by a helix bundle and a single β-strand. The two-stranded antiparallel β-sheet formed on dimerization is placed into the major groove almost perpendicular to the edges of the nucleobases so that the side chains of only one side of the β-sheet scan the DNA sequence for hydrogen bonding partners.

The minor groove can host a single β-strand. The integration host factor IHF, an *E. coli* protein used by the bacteriophage λ to integrate its genome into the bacterial chromosome, is a heterodimer with a β-hairpin extending from each monomer into the minor groove. A single β-strand of the hairpin intercalates its side chains, most importantly a conserved Pro, into the DNA. The two kinks induced by the two monomers add to a dramatic bend of almost 160° toward the IHF dimer. All six

α-helix dipoles in IHF are oriented with their N-terminus toward the ribose-phosphate backbone for favorable electrostatic interactions (see Figure 18.22).

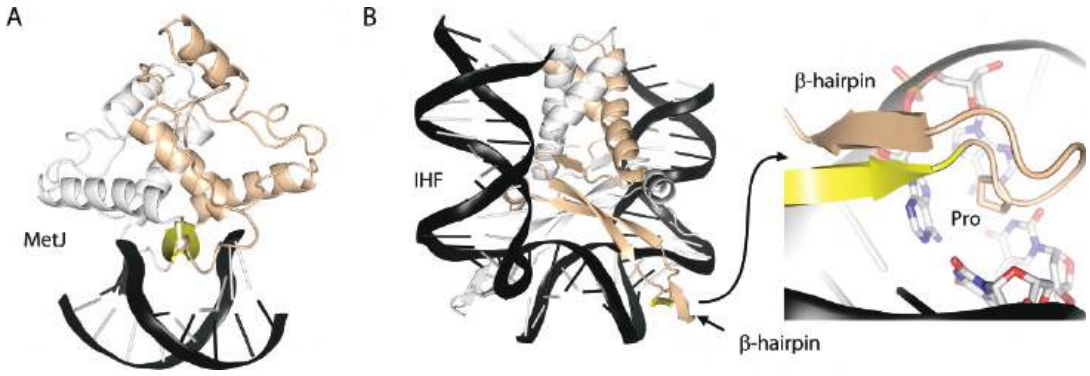


FIGURE 19.31: β-strand DNA binding motifs. A: Two β-strands (yellow) in the major groove are used for sequence-specific DNA recognition by the MetJ repressor (PDB-ID 1cma). B: The minor groove is the binding site for a single β-strand of the integration host factor (PDB-ID 1ihf). A two-stranded β-sheet inserts the edge of a single strand into the minor groove. The loop region connecting the strands has a conserved Pro residue that intercalates between two base pairs, contributing to the DNA kinking.

19.5.4.2 Small molecule binding to DNA

Binding of molecules < 1000 Da molecular mass to DNA uses the same kinds of interactions as protein-protein and protein-nucleic acid interactions, *i.e.* hydrogen bonds, ionic, and van der Waals interactions. Three main mechanisms for small molecule binding can be distinguished, *ion condensation*, *minor groove binding*, and *intercalation*.

Ion condensation is the unspecific binding of metal ions and other positively charged ions, such as spermine and spermidine to the phosphate backbone of nucleic acids. On average, 0.8–0.9 positive charges per phosphate group are associated with DNA. Metal ions, particularly Na^+ and Mg^{2+} *in vivo* counteract electrostatic repulsion of the negative charges of the phosphate backbone: the charges would otherwise be too close to allow formation of a stable structure. The large entropic loss of ion condensation is overcompensated by the favorable enthalpic contribution of strong ion pair formation. The ions are not fixed to specific binding sites: they can slide along the backbone. Ions often bind to the phosphate backbone with their hydration shells intact.

While most proteins bind to the major groove of DNA, small molecules other than metal ions either bind to the minor groove (Figure 19.32) or intercalate between base pairs (Figure 19.33). *Intercalation*, the insertion of a hydrophobic moiety between two base pairs, is driven by the hydrophobic effect, while minor groove binding is dominated by both hydrogen bonds and van der Waals interactions. Both intercalators and minor-groove binders are often positively charged, and especially minor-groove binders can engage in electrostatic interactions with the negatively charged phosphate backbone.

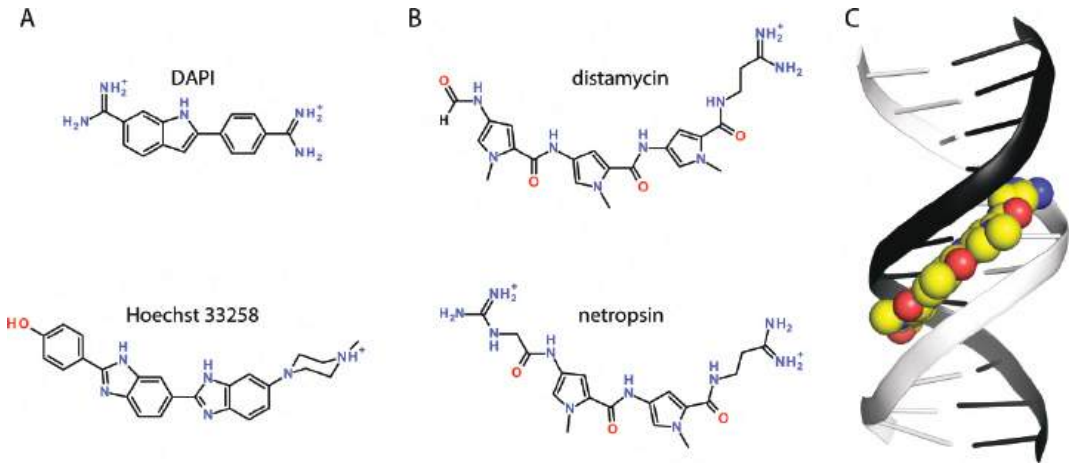


FIGURE 19.32: Molecules binding to the minor groove of DNA. A: DAPI (4′,6-diamidino-2-phenylindole) and Hoechst 33258 are used as tools in fluorescence microscopy. Both molecules bind to AT-rich sequences in double-stranded DNA, accompanied by an increase in their blue fluorescence at 460 nm. B: Netropsin and distamycin are small polymers of methyl-pyrrole connected by amide bonds. Their curvature follows that of the minor groove, as can be seen from the distamycin/DNA complex in panel C (PDB-ID 267d). All molecules have positive charges as amidinium or ammonium groups, which electrostatically interact with the ribose-phosphate backbone.

Small molecules interacting with the minor groove (Figure 19.32) have an overall crescent form that displays shape complementarity with the minor groove. They form hydrogen bonds with their NH-groups to the N3-atom of adenine and O2-atom of thymine at the sugar edge. Hydrophobic interactions are formed between the aromatic groups and the nucleobases. The water molecules normally located in the minor groove as the spine of hydration are displaced by the small molecule, so that binding is entropically favored. Minor groove binders often prefer AT-rich sequences. Binding to GC-rich sequences is disfavored because the 2-amino group of guanine bases that points into the minor groove (see Figure 19.17) sterically hinders the binding of the small molecule.

Examples of minor groove binders are DAPI and Hoechst 33258, two dyes that are used as tools for fluorescent DNA staining. The polyamides netropsin and distamycin are natural products with antibiotic and antiviral activity. Their curved shapes follow that of the DNA minor groove. Distamycin contacts up to four consecutive base pairs.

Intercalators have large aromatic planes with similar sizes to base pairs. They intercalate between adjacent base pairs without interrupting the base pairing but disrupt the stacking interactions between adjacent base pairs. The energetic penalty of the stacking disruption is overcompensated by formation of stronger stacking interactions between the intercalator and the base pairs. Binding of an intercalator is associated with a local unwinding of the DNA: each intercalated molecule lengthens the DNA by about 0.3 nm. Intercalation thus strongly unwinds and rigidifies DNA (Section 19.5.3.2), which can interfere with the action of topoisomerases to such an extent that DNA replication and transcription become impossible. Some intercalators such as the *Streptomyces*-derived anthracyclines are therefore efficient anti-cancer drugs (Figure 19.33).

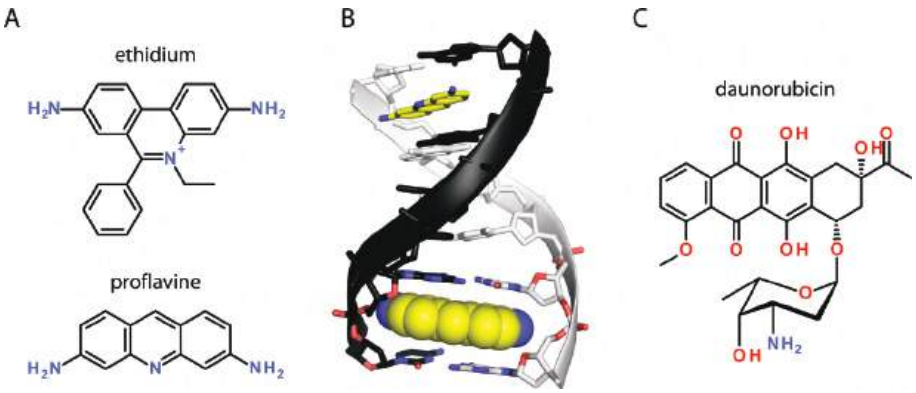


FIGURE 19.33: DNA-intercalating molecules. A: Amino groups of ethidium and proflavine hydrogen-bond to the ribose-phosphate backbone while the aromatic parts intercalate into DNA. B: Two proflavines are bound to the self-complementary DNA hexamer 5′-CGATCG-3′ (PDB-ID 3ft6) and are rendered as sticks and as space-filling model. C: Daunorubicin is a prototypic anthracycline. In addition to the intercalating aromatic part, anthracyclines contain sugar or peptidic groups that can bind to the minor groove.

19.6 RNA STRUCTURE

Many of the principles we discussed for DNA with respect to base pairing and stacking, hydrogen bonding to protein side chains, and ligand binding also apply to RNA. The main differences to DNA are in the preferred A-form of RNA double helices, the propensity of RNA to fold into complex three-dimensional shapes, and the reduced chemical stability of RNA. All of these differences can be traced back to the presence of the 2'-hydroxyl group in RNA.

The presence of the 2'-hydroxyl group in RNA strongly disfavors the B-form: in B-form RNA, a 2'-hydroxyl group in the C2'-*endo* conformation would clash with the nucleobase of the next nucleotide. Thus, the presence of the 2'-hydroxyl group changes the ribose pucker in duplexes from C2'-*endo* to C3'-*endo* (see Figure 19.9). As small as this structural change may seem, it has huge consequences on the conformation of the ribose-phosphate backbone. The torsion angles of the backbone and the ribose ring are highly correlated, and the change in ribose pucker thus also affects the overall backbone conformation. The preferred helical structure of RNA is therefore the A-form, where the 2'-hydroxyl group does not clash. As a consequence of the C3'-*endo* conformation, the 2'-hydroxyl group ends up within hydrogen-bonding distance to the O4' oxygen atom of the next ribose in the 3'-direction, which contributes to A-form stabilization. The presence of the 2'-hydroxyl disfavors the B-form to such an extent in RNA that even DNA/RNA hetero-duplexes have structurally more in common with the A-form than the B-form. The structural features of A-form RNA are similar to A-form DNA (Section 19.5.1), although the RNA double helix has much more tilted base pairs, and they are displaced away from the center of the helix toward the outside (see Figure 19.17). As a result, the minor groove of double-stranded RNA is wide and shallow, while the major groove is narrow and deep. RNA recognition by proteins thus requires other structural elements than DNA recognition (Section 19.5.4.1) and also frequently involves the 2'-hydroxyl group.

The additional hydrogen bonding capacity from the 2'-hydroxyl group helps to increase the structural variability of RNA compared to DNA. In fact, single-stranded RNA can fold into compact structures, much like a protein, where local secondary structure elements are connected by interactions between nucleotides and ribose moieties far away in sequence. In addition to the usual Watson-Crick base pairs, non-Watson-Crick, Hoogsteen, purine-purine, pyrimidine-pyrimidine, and reversed base pairs (see Section 19.1.1) that are not part of the usual structural repertoire of DNA are quite common in RNA. A number of unusual nucleobases that are post-transcriptionally introduced into RNA further increase the chemical, and thus structural, diversity of RNA compared to DNA (see Section 19.1.3).

19.6.1 RNA SECONDARY STRUCTURE AND THEIR PREDICTION

Complementary sequences in RNA molecules will pair into duplexes by standard Watson-Crick interactions just as in DNA. The RNA duplexes may be formed transiently, for example during the readout of genetic information in mRNA by tRNAs, or they can be permanent parts of RNA structure. In the latter case, duplexes serve as structural spacers, often co-axially stacked, that position the remaining nucleotides for other types of interactions. It is not uncommon for RNA duplexes to leave a few bases not engaged in Watson-Crick base pairs but located in *hairpin loops*, *internal loops*, and *bulges* (Figure 19.34). These "unpaired" nucleotides provide tertiary interactions through hydrogen bonds and stacking with other parts of the RNA that can be far away in sequence.

While larger elements are stable on their own, smaller secondary structure elements are usually embedded in larger RNA structures. Among the prominent secondary structure elements found in RNA are *stem-loop* (*hairpin*) structures, *pseudoknots*, *kink turns* (*k-turns*), *hooks*, and *S-loops* (*sarcin-ricin loops*). For hairpin formation, three or four consecutive base pairs are sufficient. As the hairpin is formed by a single RNA molecule, the complementary sequences that form the duplex (stem) necessarily have opposite directions. The loop in a hairpin sits on top of the RNA double-helix and is closed by the last base pair of the stem. Loops can have any number of nucleobases larger than three, but the one most frequently found is the *tetraloop* (Figure 19.34) with prevalent consensus sequences GNRA, UNCG, and CUUG (N being any nucleotide and R a purine). About 70% of all tetraloops in rRNAs belong to either the GNRA or the UNCG families. Tetraloops adopt defined conformations that are stabilized by hydrogen bond networks. In the most stable family, the GNRA tetraloops, the G and A nucleobases form a non-Watson-Crick base pair and the R purine hydrogen bonds to the ribose 2'-hydroxyl group of the G. The base of nucleotide N is flipped out to bind other RNA partners. In general, the loop in hairpins provides a docking site for other RNA structural elements or for protein-RNA interactions. A variant of the tetraloop is the sarcin-ricin or S-loop (Figure 19.34), a highly conserved sequence found in the 23S rRNAs of bacterial large ribosomal subunits and, in a modified form, in the 28S rRNAs of eukaryotic ribosomes. The GNRA tetraloop in the S-loop is the site of attack by the catalytic and highly toxic plant proteins α-sarcin (an endonuclease) and ricin (an N-deglycosylase), both of which inactivate the ribosome. The S-loop in addition to the hairpin has a characteristic S-shape of a few nucleotides. A bulged guanine at the center of the S-shape forms a base triple with an adjacent reverse Hoogsteen pair. A *pseudoknot* in its simplest form is generated when a single strand base pairs with a tetraloop-like structure in a hairpin. The term pseudo acknowledges that pulling at the RNA strands would unravel the structure. Two hairpin loops that contact each other to form a co-axial, pseudo-continuous RNA duplex form the *kissing loop* motif (Figure 19.34).

Similar to DNA helical junctions (Section 19.5.3.1), helical regions in RNA form two-way, three-way, and four-way junctions. Helical junctions are ubiquitous elements of RNA structure. A varying number of unpaired nucleotides directly at the junction (bulges) or as *internal loops* in the helical parts provides unsaturated valences that serve as docking sites for tertiary interactions by Watson-Crick and Hoogsteen base pairings. In contrast to the terminal loop in hairpins, the bulges and internal loops in two-way junctions are located between two helical structures and thus closed off by two base pairs.

Bulges are the sites of tight turns in the ribose-phosphate backbone. The kink turn (k-turn) is a general RNA structural motif that creates a sharp bend (90–120°) between two continuous RNA helices. One of those helices has canonical Watson-Crick base pairs, the other has non-Watson-Crick pairings. The bend occurs over three nucleotides in the bulge. Classical and reverse k-turns differ in the direction of the bend: the classical k-turn bears left upon superposition of the canonical helices and viewing from the major groove, while the reverse k-turn bends to the right (Figure 19.34).

A hook-turn is formed at sites where two strands of an RNA helix separate and continue to interact with other sequences. Such sites are often present in three-way junctions where two helices are co-axially stacked with a third helix pointing to the side at different angles. In a hook-turn, one RNA strand folds back on itself with an almost 180° reversal in direction of the ribose-phosphate backbone over just two nucleotides. The other strand remains mostly in the A-form.

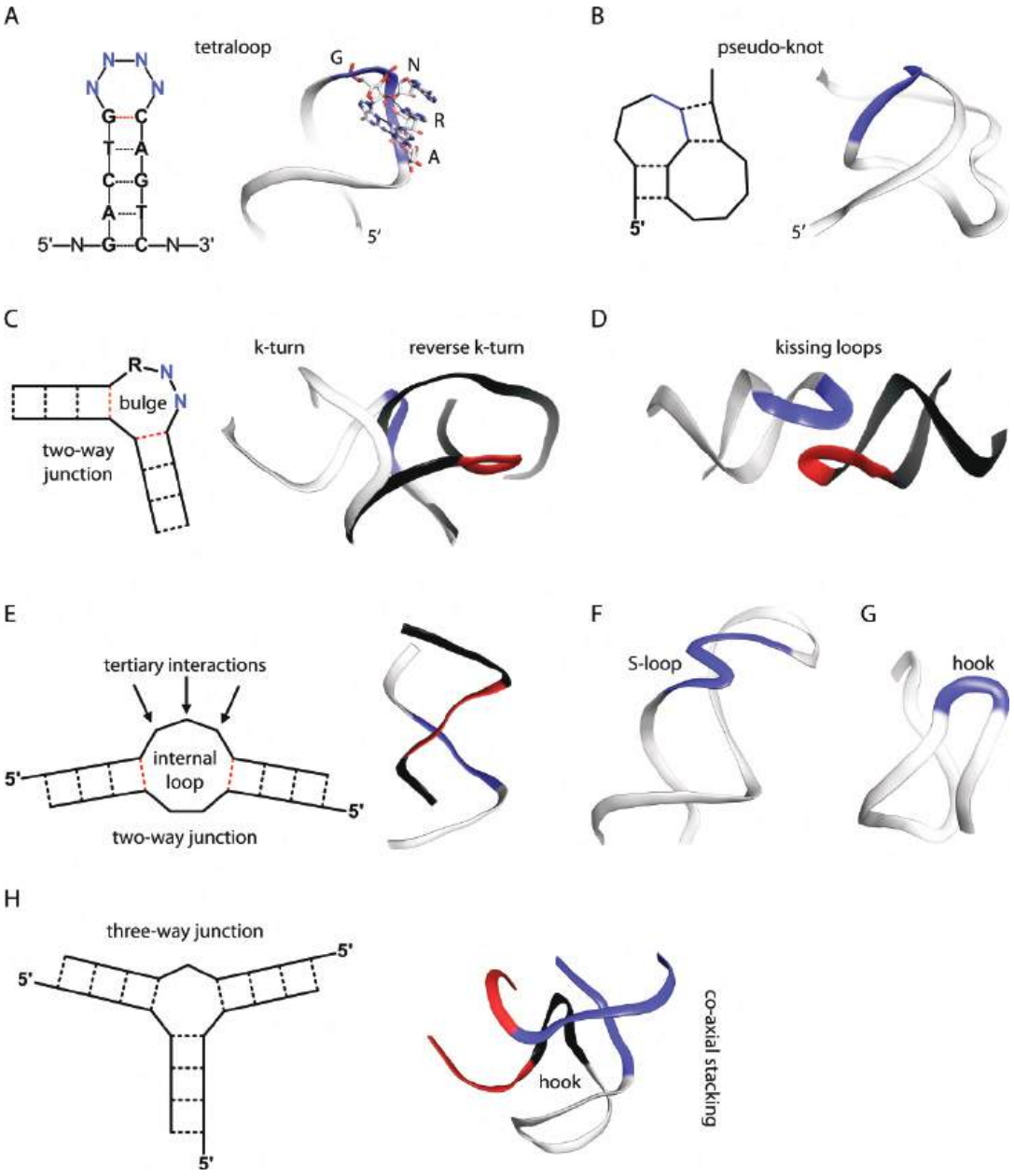


FIGURE 19.34: RNA secondary structure. Dashed lines symbolize base pairings, continuous lines follow the RNA sequence. A: The hairpin loop is formally closed by a single base pair of a standard double helix (red dashes) with a minimum of three nucleotides at its tip. The example shows a tetraloop of the GNRA family. B: A pseudo-knot is formed when a single strand of RNA contacts base pairs within a hairpin or bulge in the same molecule (from beet western yellow virus; PDB-ID 1l2x). C: Two-way junctions have unpaired regions between helical parts. A bulge is a two-way junction with one or more nucleotides in one strand "bulged out". k-turns have a three-nucleotide bulge that changes the direction of the flanking helices by 90–120° either to the left (blue bulge) or to the right (reverse k-turn, red bulge). D: Kissing loops are co-axially stacked RNA helices where the nucleotides in one loop base-pair with the nucleotides in the other loop. E: Internal loops have "unpaired" nucleotides in both strands and can be symmetric or asymmetric with the same or a different number of unpaired nucleotides in the two strands. The structure shown is of a symmetric internal loop. Bulges and interior loops are closed by two base pairs (red dashes). F: The sarcin-ricin or S-loop from *E. coli* 23S rRNA (PDB-ID483d). G: The hook turn alters the direction of the ribose-phosphate backbone by almost 180° over just two nucleotides. H: Three-way junction with a hook turn (black). Two of the helices (blue and grey) are co-axially stacked with a third helix to the side at different angles (90° in this case). Four-way junctions are also possible (see the tRNA structure in the next Section).

Secondary structures in RNA can be predicted with RNA folding algorithms, using the conserved arrangements of base pairings. The possible structures can be ranked by their relative energies. The result is a schematic plot of the predicted secondary structure with lowest energy (Figure 19.35). Different programs will calculate similar lowest energy structures, but may assign different energies to them due to different treatment of ionic strength or differences in the energy terms used for RNA folding. The most widely used energy terms for RNA folding programs are based on the *Turner1999* potential that assigns different energies to different RNA secondary structure elements, including hairpins, internal loops, bulges, multi-branch loops, and base stacking (Figure 19.34).

Base-paired regions and stem-loop structures that are close in sequence are predicted with up to 75% accuracy. Inaccuracies in the predicted structures can result from base-pair stacking in larger loops or junctions that require different energy parameters compared to standard A-RNA. The accuracy of the prediction is increased when several sequences of functionally related RNAs are available. Alignment of these sequences will identify conserved regions, which usually form the same secondary structures in different RNAs.

The larger the RNA sequence, the higher is the number of energetically similar secondary structures predicted. In these cases several conformers coexist in solution. The population of the different structures can be calculated from their energies using the *partition function* (Section 3.2). The base-pair arrangement in the lowest energy structure may only be present in a small fraction of RNA molecules.

Most algorithms start from the complete RNA sequence devoid of any local structure. Similar to the vectorial folding of proteins, *in vivo* folding of RNA will start as soon as it leaves the RNA polymerase. Co-translational RNA folding is sequential (Section 19.6.4), and will favor the formation of secondary structures between bases that are close in sequence. As a result, the calculated lowest energy arrangements using RNA folding algorithms may represent only a small fraction (< 10%) of the predominant RNA structures *in vivo*.

19.6.2 RNA TERTIARY STRUCTURE

As with proteins, tertiary structure interactions in RNA are usually established between secondary structure elements and residues far away in sequence. In this respect, pseudo-knots and kissing loops discussed in Section 19.6.1 may already count as tertiary interactions, depending on the structural and sequence context of the entire RNA molecule. The tertiary structure of tRNA may serve as a good example as it explains many principles of how RNA molecules fold. In its standard two-dimensional representation, tRNA looks like a *cloverleaf* with four base-paired regions, the anticodon stem, D-arm, T-arm, and acceptor stem. The acceptor stem contains the 5'- and 3'-ends, the others are hairpins. The base-paired regions adopt standard A-RNA helices of variable length. The anticodon stem and D-arm helices co-axially stack on top of each other, as do the helices of the acceptor stem and the T-arm. The overall tRNA structure thus comprises two helices of about 10–12 base pairs that are perpendicular to each other. Base triplets between the dihydrouridine (DHU) and pseudouridine (TΨC) loops in the D- and T-arms further stabilize the cloverleaf into a compact *L-shape*, which is typical for all tRNAs (Figure 19.35). The anticodon loop, which decodes the information in the mRNA, is located at one end of the L-shape. At the other end of the L-shape is the unpaired CCA-sequence that is esterified with an amino acid by an *aminoacyl-tRNA synthetase*. The region at the corner of the "L", formed by the DHU and TΨC loops, is termed the elbow. It serves as a binding platform for many tRNA-binding proteins, including tRNA modification and processing enzymes.

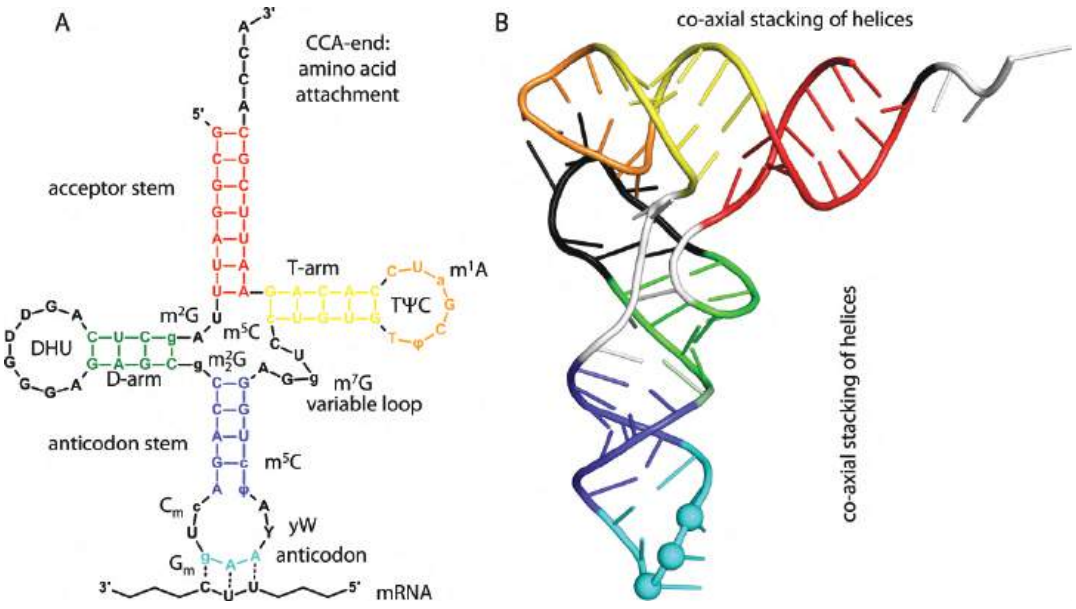


FIGURE 19.35: Structure of tRNA. A: cloverleaf representation of tRNA. B: the L-shape of tRNA is due to interaction of nucleobases from the TΨC- and DHU-loops. Two perpendicular RNA double helices are formed by several RNA pieces. The same color code is used for both representations. The sizes of the TΨC- and DHU-loops and the number of base pairs in the stems and arms are quite variable, but the total number of stacked base pairs in all tRNAs mostly ranges between 10 and 12 so that the overall shape of the molecule is conserved. The anticodon triplet is marked by three spheres. These unpaired bases stack on top of each other with their Watson-Crick edges exposed, ready to decode mRNA.

Folding of tRNAs, and of RNA molecules in general, follows the same principles as protein folding: minimize solvent-exposed hydrophobic areas and at the same time maximize polar and non-polar interactions. Minimization of hydrophobic surface exposure is achieved by base stacking. Co-axial stacking of short helices results in a longer, energetically more favorable, quasi-continuous RNA helix with uninterrupted base-pair stacking along its entire length. Indeed, co-axial stacking is the most frequently observed motif for RNA secondary structure stabilization. Stacking maximizes van der Waals interactions between nucleobases. In addition, hydrogen bonds are maximized by a variety of nucleobase interactions. Tertiary structural elements in tRNA and other RNAs include reverse Watson-Crick, Hoogsteen (Figure 19.5), and bifurcated base pairs. Often, buried hydrogen bond donors and acceptors engage in hydrogen bonds with the ribose, an interaction not observed in the standard DNA structures. Base triplets, interdigitating nucleobases, and U-turns (Figure 19.36) are further tertiary structural elements unique to RNA.

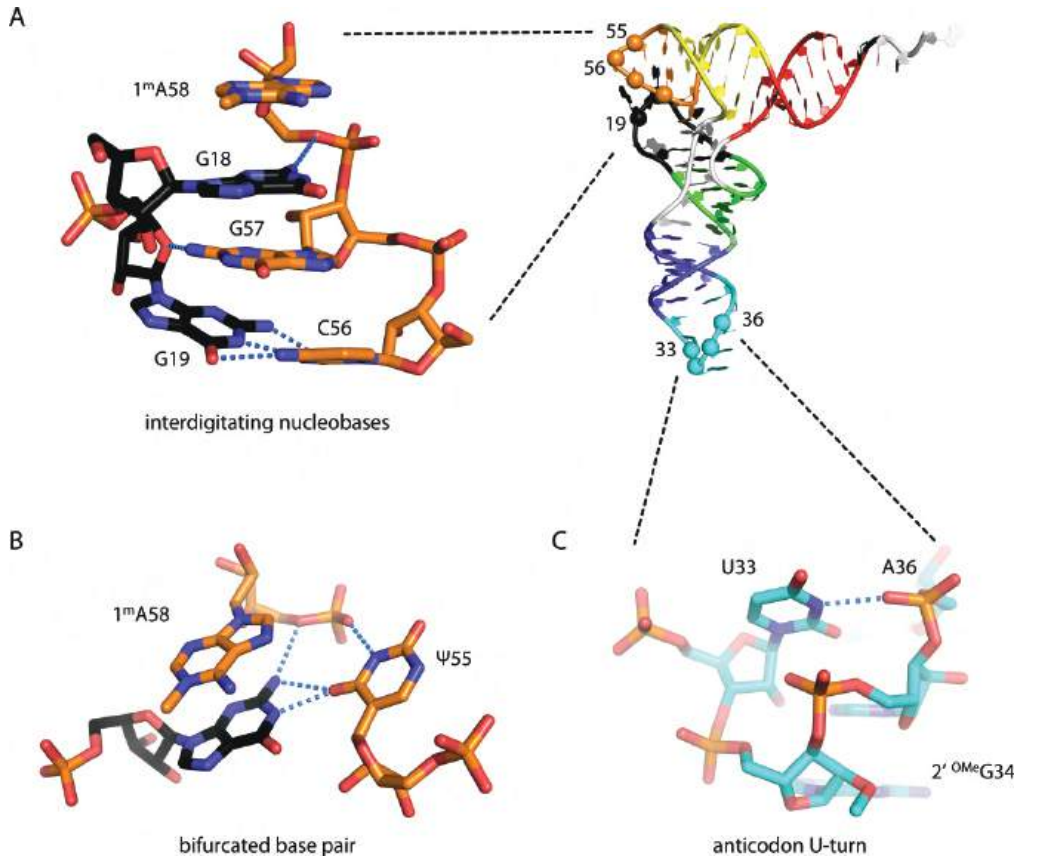


FIGURE 19.36: Nucleobase-dominated tertiary interactions in tRNA. A: The L-shape of tRNA is shown to the right with the nucleotides highlighted that are involved in the tertiary interactions. On the left, interdigitating nucleobases from the DHU (black) and TΨC (orange) loops are shown. The amino groups of guanine bases 18 and 57 hydrogen bond to oxygen atoms of the ribose-phosphate backbone (dashed blue lines). B: Bifurcated base pair with extra hydrogen bonds to the ribose-phosphate backbone from the same region as in A. C: The anticodon U-turn is closed by a hydrogen bond. These interactions are not limited to tRNA but are found commonly in RNAs.

While triple helices of DNA are artificial structures with no precedent in nature, RNA triple helices with a single strand of RNA placed in the minor groove of an A-form RNA duplex are much more prevalent. Triplex formation is possible because the minor groove in A-RNA is wider than the major groove (see Table 19.2). If the minor groove hosts just a single nucleotide from elsewhere in the RNA, a base triplet may form that is known as the A minor motif (Figure 19.37). The A minor motif is the most frequently observed RNA tertiary structure motif in RNA. It generally involves a canonical Watson-Crick "receptor" base pair that uses one or both of its 2'-hydroxyl groups for hydrogen bonding to the accepted third nucleotide. Four types (0–III) of A minor motifs have been defined, depending on the position of the 2'-hydroxyl group of the third nucleobase relative to the 2'-OH group in the receptor base pair. Type I and type II, the two most frequently observed variants, are specific for adenine bases because of their hydrogen-bonding requirements (Figure 19.37). Types 0 and III are shifted to the perimeters of the acceptor base pair and have no strict requirement for adenine.

The ribose zipper is an antiparallel contact between two different regions in a single RNA or between two RNA strands. The 2'-hydroxyl group of one strand hydrogen bonds to the nucleobase and the ribose-phosphate backbone of the second strand, and *vice versa*. Most zippers are short and contain just two consecutive nucleotides (Figure 19.37). They are frequently observed in ribozymes.

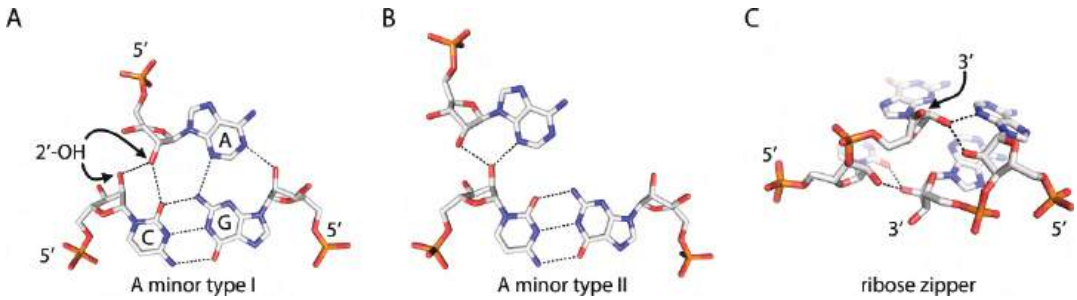


FIGURE 19.37: Ribose-dominated tertiary interactions in RNA – A minor motifs and ribose zipper. A minor motifs are ubiquitous in RNAs, specifically rRNA. The major groove is at the bottom between the 5′-phosphates. The minor groove at the top is wide enough to accommodate an extra adenine nucleotide. Types I and II both display adenine specificity. A: Type I has the adenine centrally located above the receptor base pair with hydrogen bonds to both of its 2′-hydroxyl groups. B: In type II the adenine base is shifted to one side and only has hydrogen bonds to one of the 2′-hydroxyl groups of the receptor base pair. C: The ribose zipper uses the 2′-hydroxyl group to hydrogen-bond to the opposing anti-parallel ribose-phosphate backbone and the sugar edge of the opposing nucleobases.

19.6.3 Ribozymes, riboswitches, and small molecule binding to RNA

The structural variability of RNA is substantial. It therefore comes as no surprise that RNA is able to function, quite similar to proteins, as enzymes and as binding partners for other nucleic acids, proteins, and small molecules.

The distinct tertiary structure of RNA arranges nucleobases into the catalytic sites of *ribozymes* and into the specific binding sites for ligands in *riboswitches*. Ribozymes are RNA molecules with catalytic activity, which is an interesting property for nucleic acids because prior to their discovery in the 1980s it was assumed that only proteins could act as catalysts. From an evolutionary perspective, however, RNA is older than proteins. The current *RNA-world* hypothesis speculates that during the early evolution of life all catalytic activity was brought about by RNA molecules. Until today some ribozymes work as naked RNA, including many self-splicing introns. The majority of ribozymes are parts of ribonucleoprotein complexes such as the spliceosome, the ribosome, or RNase P. Bacterial RNase P is worth mentioning here because it processes the 5′-region of all tRNA precursors. Bacterial RNase P is a complex of a small protein C5 and the large M1 RNA, which is catalytically active on its own. The catalytic framework of the M1 RNA, together with its tRNA-binding capability, was exploited in the *in vitro* evolution of flexizymes (Box 19.3). Regardless of whether ribozymes work as pure RNA or in complex with proteins, they follow the same catalytic principles as any other catalyst (Section 16): they reduce the activation energy of a reaction by preferably binding to the transition state.

BOX 19.3: EVOLUTION OF CATALYTIC AMINOACYL-tRNA SYNTHETASE RIBOZYMES - FLEXIZYMES.

Flexizymes are ribozymes that catalyze the aminoacylation of a variety of tRNAs with non-natural amino acids, hence the name “flexi”. The modified tRNAs are used for *in vitro* translation of re-programmed mRNA to introduce the non-natural amino acids into peptides and proteins (Section 18.1.5).

(*Continued*)

To evolve binding and aminoacylating activity of a flexizyme RNA, a library of 70 random nucleotide sequences is connected *via* a (constant) spacer sequence to the 5'-end of a tRNA carrying the desired anticodon. The initial idea to construct such a library and screen it for aminoacylating activity came from the crystal structure of RNase P in complex with a tRNA. The structure showed that while processing the 5'-end of the precursor tRNA, RNase P also binds to the 3'-CCA end. As the terminal adenine nucleotide of tRNA is aminoacylated for translation, the hypothesis was that some sequence added to the 5'-end of the tRNA, mimicking a tRNA "precursor", could display binding affinity for an activated amino acid and catalyze its transfer onto the CCA-end. Somewhat surprisingly, this strategy turned out to be successful. The RNA libraries are tested for aminoacylation activity using an amino acid that is activated as a cyanomethyl ester and N-terminally labeled with a biotinyl group. The chemical nature of the amino acid can be chosen widely. If the reaction proceeds, the tRNA and its covalently linked ribozyme will carry the biotinyl label. Active ribozymes can therefore be selected by binding of the biotin to streptavidin. Once a (weakly) active sequence is identified, a refined RNA library starting from this sequence is synthesized chemically. Using a mixture of nucleotides for certain positions during RNA synthesis allows focusing of the library diversification to sites that are expected to contribute to catalysis while keeping other positions known to be important for structural stability unchanged. After about 15 rounds of diversification and selection, the resulting flexizyme of about 45 nucleotide length is cleaved from the tRNA by RNase P and used to aminoacylate the tRNA for *in vitro* translation.

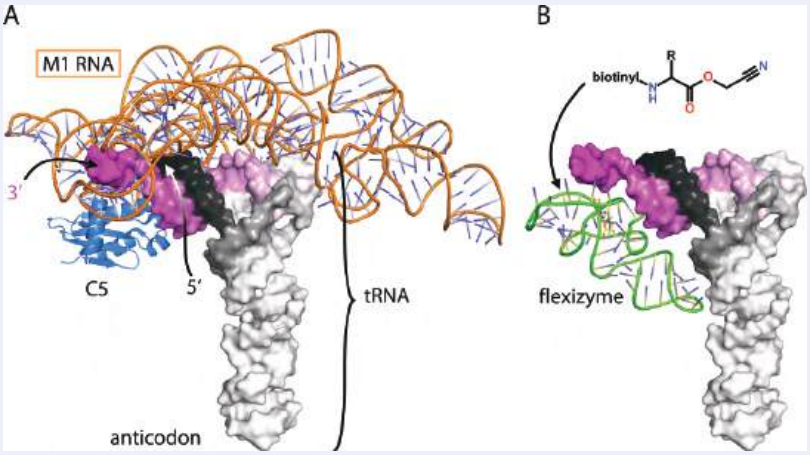


FIGURE 19.38: RNase P as a flexizyme precursor. A: Crystal structure of *Thermotoga maritima* RNase P in complex with a tRNA precursor (PDB-ID 3q1r). M1 RNA is shown as an orange ribbon and the tRNA as a surface with the 3'- and 5'-ends labeled. The 3'-CCA end of the tRNA is bound by the M1 RNA, which gave the idea to use evolved RNA for aminoacylation of tRNAs. The C5 protein (blue) increases the affinity of M1 for the tRNA, but is not necessary for *in vitro* tRNA processing or flexizyme evolution. B: The non-natural amino acid is usually activated as a cyanomethyl ester. Such esters are not exceedingly reactive, so acylation does not occur unspecifically. The biotinyl label allows selection of RNA sequences that catalyzed the aminoacylation. The structure of an evolved flexizyme is shown in green (PDB-ID 3cul), modeled onto to a tRNA for comparison with RNase P. The arrow marks the binding site of the non-natural amino acid close to the 3'-end of the tRNA.

Riboswitches are small structural domains in the 5'-untranslated regions of mRNAs that sense the concentration of a metabolite or a second messenger. Upon binding of their ligand, they undergo a conformational change (switch) that may regulate the transcription and/or the translation of the mRNA

carrying the riboswitch. Riboswitches therefore provide a feedback mechanism for the regulation of biosynthetic pathways including enzyme cofactors such as cobalamin, flavin mononucleotide (FMN), tetrahydrofolate (THF), thiamine pyrophosphate (TPP), S-adenosyl methionine (SAM), and the molybdenum cofactor Moco, as well as amino acids (Gly, Gln, Lys) and purine nucleobases. Using the SAM riboswitch as a representative, we will illustrate the principles of how small molecules bind to RNA. These principles are analogous to protein-ligand binding (Section 18.3.7) and include stacking of hydrophobic parts, van der Waals interactions, polar interactions including charge neutralization, and hydrogen bonds for specificity. Four different classes of SAM riboswitches have been described. All of them bind S-adenosyl methionine, an intermediate of methionine biosynthesis, and regulate the methionine biosynthetic pathway. The class II SAM-riboswitch consists of two short regular A-RNA helices and two loop regions (Figure 19.39). The central part of the riboswitch is a triple helix where the triplets use both the Watson-Crick and the Hoogsteen edges of their nucleobases, similar to DNA triple helices. All three helical segments stack on top of each other to form an overall linear pseudo-helical structure. The SAM binding site is in the major groove, which is narrow and deep in A-RNA, and ideally suited to host small molecule ligands. The adenine moiety of SAM intercalates between two RNA bases and participates in an AUU base triplet. The Hoogsteen edge of the adenine is recognized specifically by hydrogen bonds with one of the uridine bases. The positive charge of SAM is neutralized by nearby negative charges of the ribose-phosphate backbone.

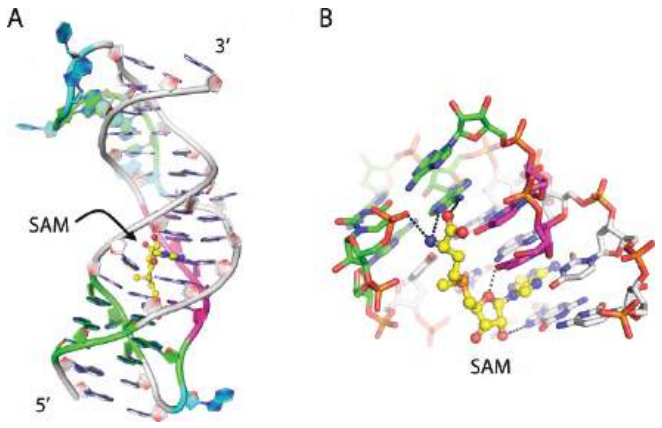


FIGURE 19.39: Riboswitches and small molecule binding to RNA. A: SAM riboswitch (PDB-ID 2qwy). Watson-Crick base pairs are shown in gray. The magenta-colored stretch forms the triple helix where SAM (yellow) binds. Green nucleotides form non-standard base pairs and cyan nucleotides have solvent-exposed nucleobases. In the absence of SAM the riboswitch adopts different structures. B: Closeup of the SAM binding site with hydrogen bonds as black dashed lines. Note the intercalation of the SAM adenine and its specific recognition by a uridine base. The positive charge on the methionine part of SAM is neutralized by a nearby phosphate group.

19.6.4 RNA FOLDING

The folding of RNA molecules into their three-dimensional structure follows similar rules to protein folding. The hydrophobic effect, which excludes water from the non-polar surfaces of the nucleobases, drives optimal stacking of base pairs, both within a single helix or between co-axially stacked helices. Buried hydrogen bond donors and acceptors are energetically unfavorable if unpaired, and hence most of the hydrogen bond donors and acceptors in RNA structures are saturated. A special driving force for RNA folding is the electrostatic repulsion of the negative charges on the phosphate backbone, many of which are solvent-exposed in the folded structure. Tertiary structure formation in RNA leads to close approach of the negatively charged phosphate backbone, making RNA folding highly dependent on the ionic strength. Cations bind to, neutralize, and connect the negative charges of the phosphate backbone, especially when the backbone is buried in the structure. The

hydrophobic core of RNA is not as extensive as in proteins, and their folding is less cooperative. Folding of RNAs can occur in a sequential fashion.

RNA folding already starts during transcription as soon as the RNA transcript leaves the RNA polymerase, leading to vectorial folding of RNA. Intermediates with stable secondary structures may form before transcription of the correct interaction sequence is finished. These intermediates are much more stable than protein folding intermediates. Thus, RNAs may get trapped in local energetic minima in conformations that may be very different from the native structure. The local minima in the RNA energy landscape are separated by high activation barriers, making intermediates difficult to resolve, and thus these minima are very effective *kinetic traps*. A progressive movement toward the global energy minimum as with proteins is therefore less likely with RNA. Instead, RNA strands in intermediate structures have to be separated and reannealed in order to allow them to fold into a more stable structure. RNA helicases are enzymes that use the energy from ATP to dissolve local RNA structure. They can act as "RNA chaperones" and catalyze structural rearrangements of misfolded RNA intermediates.

Prediction of three-dimensional RNA structure has proven more difficult than that for proteins (Section 20.3.2). Impediments to structure prediction are the extensive torsional flexibility of the RNA backbone, a much smaller structural database for RNA, and more difficult sequence alignment of RNA (4 letters) compared to proteins (20 letters). RNA structure prediction algorithms were first developed with an emphasis on maximizing the number of Watson-Crick base pairs (Section 19.6.1). However, algorithms tend to struggle with the prediction of tertiary structure interactions that strongly rely on non-classical base pairing. Coarse-grained modeling (Section 20.1.4), which substitutes pseudo-atoms for groups of atoms in a macromolecule and thus simplifies the computational problem, has been used to predict the three-dimensional structure of small (< 40 nucleotides) RNA molecules to an accuracy of 0.2–0.5 nm rmsd (root mean square distance; eq. 31.2) compared with their experimentally determined structures. More recently, machine learning algorithms have been developed that predict RNA structures with surprising accuracy, albeit not on the same accuracy level as proteins (often < 0.1 nm; see Section 20.3.2).

19.6.5 Ribonucleoprotein (RNP) complexes

RNA-protein complexes, or *ribonucleoproteins* (RNPs) for short, assemble by the same principles as DNA-protein and DNA/RNA-ligand interactions (Figure 19.40). Cations and positively charged side-chains neutralize the negative charges of the ribose-phosphate backbone and thus prevent Coulomb repulsion within the RNA. Attractive van der Waals interactions are possible between all hydrophobic surfaces, including the CH-groups of ribose (see also the TATA-box binding protein; Figure 19.28). Stacking of the planar side-chains Trp, Tyr, Phe, and His on top of or between nucleobases is a frequent interaction. Specific recognition of RNA by proteins is achieved *via* hydrogen bonding to nucleobases. Amino acid side-chains hydrogen bond to nucleobases in the same manner as in DNA-protein complexes (Figure 19.29). Hydrogen bonds may also specifically recognize the 2'-hydroxyl group in RNA.

Many ribozymes are RNPs, among them the largest known complexes, the spliceosome and the ribosome (Figure 19.40). Spliceosomes remove the non-coding introns from mRNA precursors (pre-mRNAs) that cannot self-splice. The assembly of spliceosomes is quite complex and involves intermediates with a different subunit composition. The complete spliceosome contains five U-rich small nuclear snRNAs, U1, U2, U4, U5, and U6. Dozens of proteins associate with the snRNAs, about 70 in yeast and up to 170 in higher eukaryotes. Spliceosomes are hence very protein-rich RNPs. The ribosome (Figure 19.40) is the most complicated ribozyme and RNP. Its 23S rRNA (prokaryotes) or 28S rRNA (eukaryotes) on the large subunit harbors the peptidyl transferase activity for protein biosynthesis, a reminiscence of the RNA world hypothesis. The catalytic nucleobases of the RNA are arranged as a result of RNA folding and formation of the ribonucleoprotein complex, ready to accept matching tRNAs and transfer their aminoacyl groups to the growing polypeptide chain.



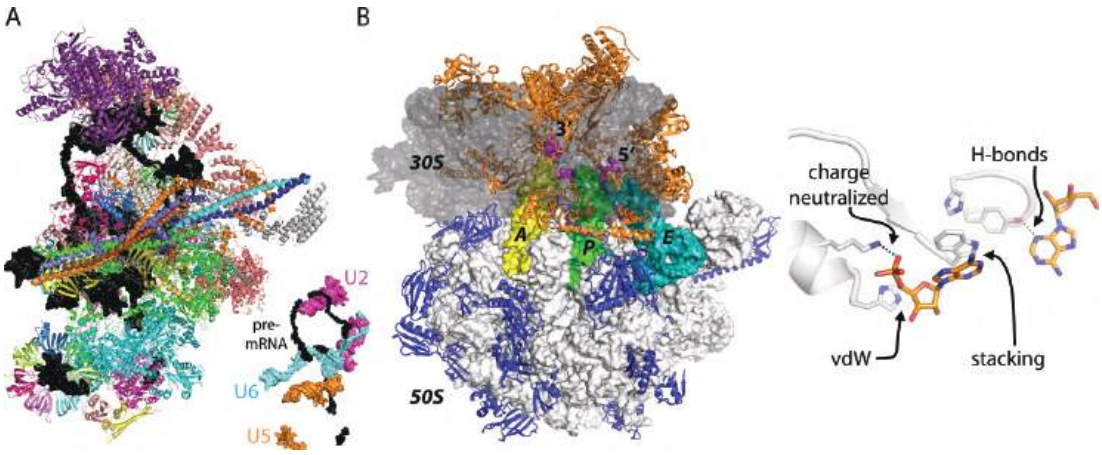


FIGURE 19.40: RNA-protein complexes. A: This cryo-EM structure of the 1.6 MDa human spliceosome (PDB-ID 5xjc) reveals 46 proteins around the RNA (7.6% by mass, shown as black surface). Removal of all proteins shows the RNA on the lower right. The pre-mRNA to be spliced (black) contacts the three snRNAs U2, U5, and U6. B: The X-ray crystal structure of the 2.2 MDa 70S ribosome of *Thermus thermophilus* (PDB-ID 4w2f) shows that ribosomes are essentially large masses of RNA (70% by mass, shown as surfaces) interspersed with < 100 small proteins. Three tRNAs are shown as differently colored surfaces contacting the mRNA (magenta) at sites A, P, and E for aminoacyl, peptidyl, and exit, respectively. The right-most panel shows some typical RNA-protein interactions.

Among the most useful discoveries in the field of RNPs are the CRISPR-Cas systems. CRISPR-Cas RNPs are part of an inheritable immune system that enables bacteria to defend themselves against invasive DNA from viruses and plasmids. Bacteria integrate short non-self DNA sequences from previous infections into a locus called CRISPR (Clustered Regularly Interspersed Short Palindromic Repeats). The presence of this foreign DNA allows the bacteria to recognize an infection with the same or similar sequences, and then destroy the invading DNA with the help of the nuclease Cas (CRISPR-associated). Recognition of foreign DNA is achieved by transcribing the DNA stored in the CRISPR locus into a precursor CRISPR RNA (pre-crRNA) that is further processed into two RNA molecules, crRNA and trans-activating CRISPR RNA (tracrRNA). The two RNAs and Cas assemble into the mature RNP, which scans foreign DNA for target sites matching the sequence of the crRNA. A conserved protospacer flanking sequence (PFS) that is specific for a certain Cas enzyme also must be present in the target DNA. Once a PFS is recognized, Cas unwinds the double-stranded DNA and thus allows for probing its interaction with the tracrRNA. After hybridization of foreign DNA to the tracrRNA, the nuclease cleaves the intrusive DNA on both strands and thus abolishes the infection. This sequence of events is common to the five different CRISPR systems currently known; different systems deviate from the common principle in different ways.

The simplest of these systems, CRISPR-Cas9 (Figure 19.41), has recently been developed into a powerful gene editing method (Box 19.4) that is now being used to treat genetically inherited diseases and various cancers by replacing or repairing genes. Also, there is hope that this technique will enable the excision of entire retroviral genomes from chromosomes in the near future.

BOX 19.4: GENE EDITING BY CRISPR-Cas SYSTEMS.

Similar to zinc finger nucleases (Box 19.2), the CRISPR-Cas system can be used for gene editing with exquisite sequence specificity. For this application, the system is simplified from its natural antecedent by joining the tracrRNA and crRNA into a single RNA, called the single guide RNA (sgRNA). The PFS is now called PAM (protospacer adjacent motif), and its choice depends on the Cas variant used. The simplest of these systems is CRISPR-Cas9,

(*Continued*)



which accepts the NGG sequence as PAM. The 160 kDa nuclease Cas9 forms a complex with the sgRNA, whose sequence is matched to that of a target DNA connected by a spacer to the NGG PAM. Typical lengths of the target DNA-recognizing sequences in sgRNAs are 20 nucleotides, which normally provides sufficient specificity and affinity to keep the Cas9 from cleaving DNA unspecifically at other than the intended location (Figure 19.41).

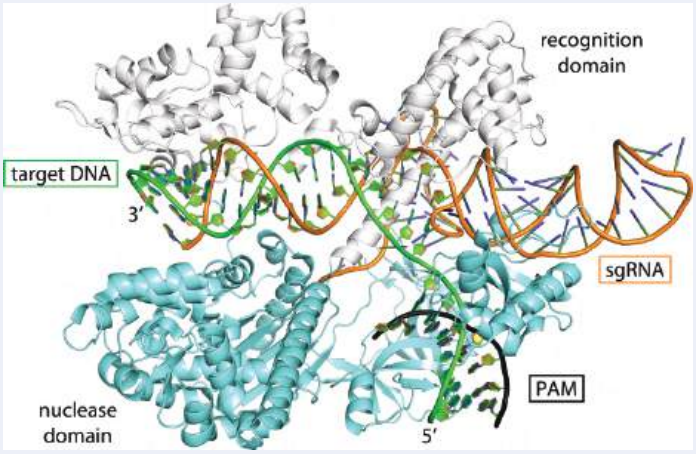


FIGURE 19.41: CRISPR-Cas9 RNP complex. One of the smallest complexes is that from *Staphylococcus aureus* (PDB-ID 5axw). The single guide RNA (orange) binds to the Cas9 recognition domain (grey) and forms a hetero-duplex with the target DNA (green). For the sake of structure determination, the DNA was elongated to include the sequence complementary to the PAM (black). The PAM is recognized in a sequence-specific manner by the nuclease domain (blue).

The designed Cas9-sgRNA complex is packaged into liposomes (Box 18.5) or adeno-associated viruses (AAVs) for delivery into cells. The initial recognition of the target DNA in the cell happens *via* the PAM. Once the target DNA has formed a hetero-duplex with the sgRNA, a double-strand break is introduced. The break is repaired by the cell in two different ways, either by non-homologous end joining (NHEJ) or by homology-directed repair (HDR). NHEJ is an error-prone repair mechanism where the ends of DNA are simply rejoined, leading to insertions, deletions, and/or frame shifts. Unless the aim is to inactivate a certain gene, this outcome is usually undesired. The more precise HDR relies on homologous recombination using an exogenous DNA template with the desired "healthy" sequence. This DNA template is co-delivered along with the Cas9-sgRNA RNP, and, surprisingly enough, the cell uses the template to repair the double-strand break. The size of the DNA template, when packaged in AAVs is < 5kb.

QUESTIONS

19.1 Draw the structure of 2'-deoxy-GTP at physiological pH including all stereo centers. Assign two different pK_A-values (about 2 and 7) to the phosphate oxygen atoms. How many charges does dGTP carry under physiological conditions? Explain the difference between a phosphoric acid anhydride and a phosphoester from this structure.

19.2 Draw the structure of deprotonated 1-methyl-adenosine. Can you suggest why the protonated form is much more stable? What about 3-methyl-cytosine?



19.3 Free riboses are hemi-acetals and can open to an aldehyde. The closed furanose is much more stable than the open aldehyde. Upon closure of the aldehyde, two anomeric configurations can form that have different chirality at C1'. Draw the aldehyde form of ribose and explain how the two anomers can be formed upon ring closure.

19.4 Bromination of guanosine at C8 favors the *syn*-conformation. What could be the explanation?

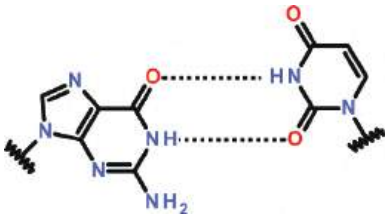
19.5 At room temperature and neutral pH, RNA spontaneously hydrolyzes about 100 times more quickly than DNA. The rate of RNA hydrolysis is considerably increased in alkaline solutions. What might be the reason and which chemical reaction is responsible for RNA instability?

19.6 Draw the G-C Hoogsteen base pair. Why is a protonated cytosine required?

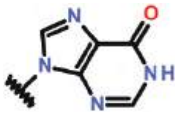
19.7 The A-T base pair is more stable than the A-U base pair. Can you explain?

19.8 Draw the enol forms of G and T. Why can they not base-pair with C and A, respectively? What if all nucleobases are converted to their tautomers?

19.9 The G-U Watson-Crick base pair is shown. Why can 2-thiouracil not base-pair with G but only with A?



19.10 Inosine (shown) is often present at the first position of the anticodon of tRNA (the wobble position) where it recognizes U, C, and A. Draw the three Watson-Crick base pairs for inosine.



19.11 The codon AUA codes for isoleucine, while AUG codes for methionine. How can inosine distinguish between the two codons?

19.12 The RNA sequence 5'-GGCCGAAAGGCC-3' can form secondary structure as a monomer and as a dimer. Draw the schematic structures with base pairs and identify the RNA secondary structure elements.

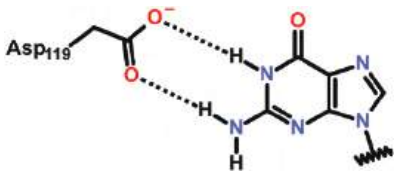
19.13 A plasmid of 1000 bp has a superhelical density of $\sigma = -0.05$. Calculate Lk, Lk^0, and ΔLk. How does σ change after 5 catalytic cycles of DNA supercoiling by gyrase?

19.14 Why is DNA relaxation an exergonic reaction although the DNA backbone needs to be cleaved?

19.15 What is the overall bend of a DNA where 4 poly-A tracts are each spaced by 5 random base pairs?



19.16 Acidic side chains can form a bi-dentate hydrogen bond to the Watson-Crick edge of guanine bases. An Asp side chain binds to the guanine in Ras, as shown in the figure.



A variant of Ras, Asp119Asn, shows much weaker affinity for guanine nucleotides but strongly binds to xanthine nucleotides. Xanthine differs from guanine by substitution of an oxygen atom for the exocyclic amino group. How could the switch in nucleotide specificity be explained?

REFERENCES

Bryan, T. M. (2020) G-quadruplexes at telomeres: Friend or foe? *Molecules* **25**(16): 1–22.
· review on G-quadruplexes at telomers

Devi, G., Zhou, Y., Zhong, Z., Toh, D. F. and Chen, G. (2015) RNA triplexes: From structural principles to biological and biotech applications. *Wiley Interdiscip Rev RNA* **6**(1): 111–128.
· review on RNA triplexes

Egli, M. and Saenger, W. (1988) *Principles of Nucleic Acid Structure*. Springer.

Blackburn, G. M. (2006) *Nucleic Acids in Chemistry and Biology*. Cambridge, RSC Pub.

Sinden, R. R. (1994) *DNA Structure and Function*. San Diego, Academic Press.
· comprehensive treatment of all aspects of nucleic acid chemistry and structure

Gilbert, S. D., Rambo, R. P., Van Tyne, D. and Batey, R. T. (2008) Structure of the SAM-II riboswitch bound to S-adenosylmethionine. *Nat Struct Mol Biol* **15**(2): 177–182.
· structure of the SAM riboswitch bound to SAM

Gotoh, O. and Tagashira, Y. (1981) Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles. *Biopolymers* **20**(5): 1033–1042.
· base stacking

Haran, T. E. and Mohanty, U. (2009) The unique structure of A-tracts and intrinsic DNA bending. *Q Rev Biophys* **42**(1): 41–81.
· review on DNA bending

Kouzine, F. and Levens, D. (2007) Supercoil-driven DNA structures regulate genetic transactions. *Front Biosci* **12**: 4409–4423.
· supercoiling *in vivo*

Kouzine, F., Gupta, A., Baranello, L., Wojtowicz, D., Ben-Aissa, K., …, Levens, D. (2013) Transcription-dependent dynamic supercoiling is a short-range genomic force. *Nat Struct Mol Biol* **20**(3): 396–403.
· supercoiling *in vivo*

Kouzine, F., Sanford, S., Elisha-Feil, Z. and Levens, D. (2008) The functional response of upstream DNA to dynamic supercoiling *in vivo*. *Nat Struct Mol Biol* **15**(2): 146–154.
· supercoiling *in vivo*

Krishna, S. S., Majumdar, I. and Grishin, N. V. (2003) Structural classification of zinc fingers: Survey and summary. *Nucleic Acids Res* **31**(2): 532–550.
· review on zinc fingers

Leontis, N. B. and Westhof, E. (2001) Geometric nomenclature and classification of RNA base pairs. *RNA* **7**(4): 499–512.

Introduction of the Watson-Crick edge, the Hoogsteen edge, and the sugar edge to better describe nucleobase interactions

Lescoute, A. and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA* **12**(1): 83–93.
· Three families of three-way junctions are discussed



Ornstein, R. L. and Rein, R. (1978) An optimized potential function for the calculation of nucleic acid interaction energies I. base stacking. *Biopolymers* **17**(10): 2341–2360.

· Stacking energies between complementary pairs of dinucleotide dimers

Pan, C., Potratz, J. P., Cannon, B., Simpson, Z. B., Ziehr, J. L., Tijerina, P. and Russell, R. (2014) DEAD-box helicase proteins disrupt RNA tertiary structure through helix capture. *PLoS Biol* **12**(10): e1001981.
· RNA tertiary structure disruption by an RNA helicase

Potratz, J. P., Del Campo, M., Wolf, R. Z., Lambowitz, A. M. and Russell, R. (2011) ATP-dependent roles of the DEAD-box protein Mss116p in group II intron splicing *in vitro* and *in vivo*. *J Mol Biol* **411**(3): 661–679.
· ATP-dependent unwinding to help RNA folding

Rhodes, D. and Lipps, H. J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucl Acids Res* **43**(17): 8627–8637.
· review on G-quadruplexes

Rich, A. and Zhang, S. (2003) Timeline: Z-DNA: The long road to biological function. *Nat Rev Genet* **4**(7): 566–572.
· review on Z-DNA

Sokoloski, J. E., Godfrey, S. A., Dombrowski, S. E. and Bevilacqua, P. C. (2011) Prevalence of *syn* nucleobases in the active sites of functional RNAs. *RNA* **17**(10): 1775–1787.
· prevalence of *syn* nucleobases in the active sites of functional RNAs

Vicens, Q. and Kieft, J. S. (2022) Thoughts on how to think (and talk) about RNA structure. *PNAS* **119**(17): e2112677119.

Highlights that two-dimensional plots of RNA secondary structure are overstating Watson-Crick base pairing and understating the importance of the non-Watson-Crick interactions.

Zhang, J. and Ferré-D'Amaré, A. R. (2016) The tRNA elbow in structure, recognition and evolution. *Life* **6**(1): 1–11.
· Discusses the elbow at the kink in the L-shape of tRNA.

Online resources

modomics.genesilico.pl/modifications
database of RNA modifications
mfrlab.org/interrna
· database of tertiary RNA interactions



20

Macromolecular modeling

Bioinformatics became a research area in the late 1970s when computer power started to increase and more and more data could be handled and stored. Statistical methods were developed to analyze data that were then stored and curated in databases. The first databases stored sequences of proteins and DNA, determined with classical and time-consuming (semi)-manual sequencing techniques. With the advent of high throughput sequencing techniques, often combined with mass spectrometry, sequences of entire genomes, transcriptomes (the sum of all RNAs), and proteomes (all of the proteins in the cell) became accessible. These methods generate large datasets, and these biological databases continue to be the largest that we have. Genome annotation remains one of the most fruitful activities in bioinformatics, identifying and ascribing likely biological functions to genes and gene products. However, it quickly became clear that the number of genes in a genome is too small to explain the large variety of proteins and functions in the cell. Recent databases contain content from other "*omics*" fields, such as "metabolomics" (the pool of metabolites and their interconversion in cells), "glycomics" (the glycosylation patterns in organisms), and "interactomics" (the sum of all interactions between molecules in the cell). In line with the term *genome* for the genetic material, the new datasets are aptly named *transcriptome*, *proteome*, *metabolome*, *glycome*, and *interactome*, respectively. A survey of the methods used in bioinformatics and the types of information gained from them is given for example at www.expasy.org and www.click2drug.org. Among the first applications of computational biology were sequence analysis and molecular modeling. Sequence compositional analysis and sequence alignments for proteins and nucleic acids allow secondary structure predictions and the generation of phylogenetic trees, as discussed in Sections 18.2 and 19.3, respectively. Molecular modeling uses force fields and energy terms to describe molecular structure and motions (Section 20.1). A more recent development in computational biology is the routine use of neural networks machine learning (a subset of "*artificial intelligence*", or "AI" algorithms) for the analysis of large datasets and making predictions from them (Section 20.2). Protein folding *in silico* especially has benefited from this development (Section 20.3).

20.1 MOLECULAR MECHANICS AND DYNAMICS

Molecular modeling is the umbrella term for numerically representing molecules and simulating their structures as a function of time. Such simulations are approximate, and their results must be



validated by comparison of the results with experimental data. Molecular modeling is applied in many areas of research, from finding the global energy minimum of small molecule conformations *via* folding of entire proteins to the study of supramolecular complexes as large as the ribosome.

Molecular modeling uses the laws of molecular mechanics (MM) and/or quantum mechanics (QM). MM and *molecular dynamics* (MD) calculations adequately describe many aspects of molecular conformation, conformational changes, and rigid-body movements, but some molecular properties require the more detailed QM treatment. Examples include processes that involve drastic changes in electronic structure, such as fission and formation of chemical bonds when modeling catalysis, and the conformational description of small molecules using molecular orbitals. While QM treatment of small-molecule geometry is becoming routine practice in applications such as crystallographic structure refinement, the problem with QM methods is that they are computationally very demanding. Pure QM calculations on systems with more than 1000 atoms are so time-consuming that is often necessary to limit the QM treatment to a small part of the system, *e.g.* a ligand in a protein-ligand complex, while the rest of the system is described by laws of classical mechanics MM/MD methods.

A central element in MD simulations is the *force field* that expresses the physico-chemical information that we have about molecules as a sum of energy terms in a so-called *potential energy function*. With the force field, the potential energy of a system (a lipid bilayer, a hydrated protein, a protein-nucleic acid complex, *etc.*) can be calculated. Thus, a crucial aspect of MM/MD methods is the choice and parametrization of the force field because this will have a strong effect on the calculated potential energy. The energy terms in a force field are empirical but it is *a priori* not known which ones dominate the potential energy, so the terms need to be weighed against each other. Force fields also cannot take quantum effects into account, which is unfortunate because QM calculations currently result in the most accurate descriptions of molecules. Therefore, the energy terms constituting the force fields are weighted against each other in such a way that the potential energy calculated agrees with QM calculations on small molecule reference compounds. The assumption is that the QM information from small molecules is transferable to large molecules that cannot be treated by QM because of excessive computation times. Among the force fields that have been developed to study the structure and dynamics of proteins and nucleic acids are CHARMM, AMBER, GROMOS, MOLOC, and ROSETTA. In the following, we will take a look at force fields and how they are used in MM/MD studies.

20.1.1 Force fields

There are two main contributors to the potential energy of a molecule, interactions *via* covalent bonds and non-covalent interactions through space (Chapter 17):

$$E_{\text{molecule}} = \sum\left(E_{\text{covalent}}\right) + \sum\left(E_{\text{non-covalent}}\right) \qquad \text{eq. 20.1}$$

Covalent interactions are described by parameters such as bond distances, 1,3-angles, and torsion (1,4-) angles plus additional *improper dihedral* restraints that maintain flatness of planes, *e.g.* of peptide bonds and nucleobases, and keep the chirality of stereocenters. The overall non-covalent interactions are typically implemented as the sum of many individual energy terms (Section 17.2.2) that are weighted against each other to reflect their relative importance. Non-covalent energy terms are chiefly the van der Waals interactions and all interactions between permanent dipoles, including hydrogen bonds. The ROSETTA force field used in the prediction of protein structures, for example, contains specific energy terms for atom solvation in the folded state and amino acid residue desolvation upon folding. It divides hydrogen bonds in short-range and long-range, and whether they are formed between backbone or side-chain atoms. There can even be different energy terms for hydrogen bonds depending on the hybridization state of the contributing atoms. The possibilities of shaping the force field by adding specialized energy terms appears endless. However, for a molecule of N atoms, each additional energy term increases the number of energies that have to be calculated by $N{\cdot}(N{-}1)$, which



slows down the calculations. The force field can be simplified again by incorporating known facts about the molecule under study, such as the presence of cofactors, covalent modifications of residues, or conserved protein regions that should not move during the simulation. These simplifications reduce the number of degrees of freedom of the system, speeding up the calculations. The ROSETTA force field, to keep the example, is simplified by inclusion of empirical information on preferred torsion angles according to the Ramachandran potential for the protein backbone (Figure 18.21 in Section 18.2), and on preferred rotamers for the side chains (Figure 18.5 in Section 18.1.3). This "fine-tuning" of force fields highlights an important aspect – they can incorporate prior knowledge and experimental data from many different biophysical and structural sources to increase the confidence in the energy minimization. This approach is termed *hybrid methods*, and is now applied routinely to large and flexible macromolecular complexes for which one particular method yields only incomplete data (Box 20.1).

BOX 20.1: HYBRID OR INTEGRATIVE METHODS.

Hybrid methods integrate the information from complementary structural methods to increase the data-to-parameter ratio and to arrive at a comprehensive structural description of the molecule of interest. Hybrid methods may combine geometric and distance restraints from nuclear magnetic resonance (NMR; Section 22.1), X-ray diffraction (Chapter 24), Förster resonance energy transfer (FRET; Section 21.5.10), and electron paramagnetic resonance (EPR; Section 22.2). These methods are discussed in Part IV. The experimental data from each method are represented by additional energy terms, integrated into a single force field, that is then used for energy minimization and structure calculation. This approach is particularly useful for combining short-range distance restraints from NMR with long-range distance restraints from EPR (Box 22.9), or restraints from NMR and FRET, or restraints from EPR and FRET. The hybrid concept also finds applications in the structural modeling of large macromolecular complexes: molecular envelopes from small angle X-ray scattering (SAXS, Section 23.2.4) or cryo-electron microscopy (EM) maps (Section 26.5) can be filled by rigid-body modeling of NMR and crystal structures to gain insight into the complex interfaces, and conformational changes monitored by scattering, fluorescence, or NMR can be compared to the models from MD simulations. The accuracy of these envelope fitting and modeling procedures can be validated if biochemical information about protein-protein interactions is available by mutagenesis or by crosslinking data from mass spectrometry. Using a combination of EM and crosslinking mass spectrometry data, together with docking of X-ray crystal structures and homology modeling, the structure of a large part of the nuclear pore complex could be visualized. Bui *et al.* (2013). Cell 155(6): 1233–1243.

20.1.2 Energy minimization

Energy minimization is the search for the minimum of the potential energy function by gradually changing the Cartesian coordinates r_i of the atoms of the system under investigation. Calculus of analytical functions tells us that in order to find their minimum, the first derivative should be zero. The force F is the first derivative of the potential energy E with respect to the atom coordinates, hence at the minimum of the energy function, there are no net forces acting on the atoms.

$$-\frac{\mathrm{d}E}{\mathrm{d}\boldsymbol{r}} = -\nabla E = F = 0 \qquad \text{eq. 20.2}$$



Here, $\boldsymbol{r}$ in bold face is the vector describing the atomic positions r_i and ∇ is the Nabla operator signifying the vector differential, or total differential of the energy function for all variables r_i. Energy minimization is usually carried out in Cartesian space, so a set of N atoms is described by $3N$ variables. In contrast to analytical functions where the exact minimum can be found using calculus, molecular motion follows force fields that are too complex to solve analytically. The derivative of E has to be calculated numerically. In numerical methods, gradual changes of the atomic positions are introduced, E is recalculated, and the new set of coordinates is only kept if the energy is lower than the starting energy. After a few thousand iterations, the procedure converges at a conformation of low energy.

Several protocols/strategies/approaches are in use for energy minimization, among them the *steepest descent*, the *conjugate gradient*, and the *Newton-Raphson* algorithm. All of them use derivatives of the potential energy function to calculate the minimum, but they differ in their paths to approach the minimum.

The steepest descent method calculates the force as the first derivative of E. Atoms are then moved in the direction of the force. Each movement is perpendicular in direction to the previous step, by definition. The decision how far the atoms are moved can be made based on the magnitude of the force, reflecting the steepness of the energy surface. Alternatively, the atoms can be moved by an arbitrary step size. Flat energy surfaces warrant larger step sizes while sampling of narrow valleys with steep gradients needs smaller step sizes, but the shape of the surface is not known *a priori*. At the new atomic positions, the energy and its derivative are recalculated. The derivative provides the new force and its new direction perpendicular to the previous step, and so on. Each energy is compared to the energy of the previous conformation. If the energy decreases, the step size is increased by a factor and the procedure repeated until the energy stops decreasing. If, after some iterations, the energy increases again, the molecule has leapt over an energy minimum, moving uphill on the energy surface. In this case, the system, *i.e.* the ensemble of atoms, is moved back into the direction of the minimum by a *line search* where a line is drawn between the last three sampled points (representing molecular conformations) on the energy surface. As the energy has increased in the second step, the minimum must be between the first and the third point. The algorithm must then decide on where to move the atoms between the first and third energy level to get the system back on track. A computationally efficient way is to approximate the (curved) line on the energy surface by a quadratic function that is fit to the three energy levels. Its minimum, which is close to but not necessarily identical with the second point, is used as the new starting coordinates. While it is a fast algorithm, the steepest descent tends to overshoot in narrow energy valleys and tends to become trapped in local energy minima, where it oscillates about the energy minimum.

The steepest descent method moves atoms perpendicular to the previous step, which may not always be the best choice as it may lock the system oscillating around a local energy minimum. The conjugate gradient method avoids the oscillating tendencies of steepest descent by a different choice of direction. The first iteration is the same as in the steepest descent method, but starting from the second iteration the direction of the atom movement is not necessarily perpendicular to the previous step. Instead, the new direction vector $\boldsymbol{d}_{\text{new}}$ is a conjugate between the gradient $\boldsymbol{g}_{\text{new}}$ at that starting point and the direction vector $\boldsymbol{d}_{\text{previous}}$ of the previous step.

$$\boldsymbol{d}_{\text{new}} = -\boldsymbol{g}_{\text{new}} + \gamma_{\text{new}} \cdot \boldsymbol{d}_{\text{previous}} \qquad \text{eq. 20.3}$$

This means that the direction of the second step incorporates information from the first step, which is an advantage over the steepest descent method as in practice it leads to convergence more quickly. The scaling constant γ_{new} is a scalar calculated from the ratio of the following scalar products:

$$\gamma_{\text{new}} = \frac{\boldsymbol{g}_{\text{new}} \cdot \boldsymbol{g}_{\text{new}}}{\boldsymbol{g}_{\text{previous}} \cdot \boldsymbol{g}_{\text{previous}}} \qquad \text{eq. 20.4}$$



γ_{new} in eq. 20.4 corresponds to the *Fletcher-Reeves algorithm*, but other algorithms define γ_{new} differently. Just as with the steepest gradient method, step sizes for the conjugate gradient method are either determined by a line search or are set arbitrarily.

In contrast to the steepest descent and conjugate gradient method, the Newton-Raphson algorithm uses not only the first but also the second derivative of the potential energy function to locate the energy minimum. The second derivative provides information on the curvature of the energy surface, hence it tells us whether the energy minimum is flat or pointed. The Newton-Raphson algorithm approximates the potential energy function by a Taylor series (Section 31.14.1) around the point at coordinates r_i:

$$E(\boldsymbol{r}) = E(r_i) + E'(r_i)\cdot(r - r_i) + \frac{E''(r_i)}{2}\cdot(r - r_i)^2 + \cdots \qquad \text{eq. 20.5}$$

Just as with the line search procedure for the steepest descent method, the energy function is approximated by a quadratic function and the expansion is therefore stopped at the quadratic term. The first and second derivatives of the potential energy function in eq. 20.5 are

$$E'(\boldsymbol{r}) = E'(r_i) + r\cdot E''(r_i) - r_i\cdot E''(r_i) \qquad \text{eq. 20.6}$$

and

$$E''(\boldsymbol{r}) = E''(r_i) = \text{const.} \qquad \text{eq. 20.7}$$

When fitting the energy states with a quadratic function, the second derivative, or curvature of the energy function, is a constant. The energy is minimal for the condition $E'(\boldsymbol{r}) = 0$, and hence it follows from eq. 20.6

$$E'(\boldsymbol{r}) = 0 \Rightarrow r = r_i - \frac{E'(r_i)}{E''(r_i)} \qquad \text{eq. 20.8}$$

The advantage of this algorithm over the steepest descent and the conjugate gradient methods is that the "flatness" of the energy function is taken into account when estimating the step size. For a multidimensional function, the second derivative $E''(\boldsymbol{r})$ transforms into a huge *Hessian* matrix that must be inverted at considerable computational cost to solve eq. 20.8 for r. The Newton-Raphson algorithm is powerful and for a quadratic potential may find the energy minimum in a single step, regardless of the starting value. Often, potential energy functions formulated for molecules are more complex, and minimization then requires several steps. At each of them, matrix inversion has to be performed, which limits the size of the systems that can be handled to a few hundred atoms. Variants of the Newton-Raphson algorithm seek to avoid the time-demanding matrix inversion to speed up the calculations. For example, the same Hessian matrix can be used unaltered for a number of steps and only $E'(\boldsymbol{r})$ is recalculated at each step.

20.1.3 Simulated annealing

The conformational energy landscape of a macromolecule is rugged with many local minima, separated by energy barriers (Figure 18.55). Energy minimization methods therefore run the risk of getting trapped in a local energy minimum. To approach the global energy minimum, the ensemble of atoms (a single molecule or a molecular complex) needs to overcome the energy barrier separating one energy minimum from the next. This becomes possible by *simulated annealing*. In this procedure, the macromolecule is "heated" *in silico* to the desired starting temperature of several thousand K. Heating *in silico* is equivalent to assigning large random velocities to the atoms so the system can adopt high-energy conformations to overcome energy barriers separating low-energy conformational states (local energy minima). Other than increasing the likelihood of overcoming



local energy barriers, the temperature in simulated annealing has no physical meaning. Because of the heating step that introduces a large amount of energy, the *radius of convergence* is larger for simulated annealing compared to other energy minimization methods, *i.e.* the system may start in a conformation far away from its global energetic minimum and still converge to the true structure. After heating the system the next step is slow cooling, also known as *annealing*. The progressively lower temperatures during annealing allow the atoms to move more slowly and thus to engage in more favorable interactions with each other compared to previous conformations at higher temperature. In other words, with lower temperatures lower energy conformations become more probable, in accordance with the Boltzmann distribution (eq. 21.7). The annealing process is guided by force fields and experimental data to avoid divergence of the structure and to ensure a chemically sensible result. To find the global energy minimum with simulated annealing, in principle an infinitely large number of infinitesimally small temperature steps would be required. Simulated annealing is therefore a compromise between computation time and accuracy, and it is always uncertain whether the global energy minimum has been found. In practice, this uncertainty is addressed by performing several cycles of simulated annealing at different starting temperatures and with different cooling rates. If the results are identical across different runs, it is quite likely that the global minimum has been reached. Simulated annealing may also return a series of different structures of similar energy, a typical result for NMR structure calculations that lends insight into the dynamics of the molecule. Apart from NMR, simulated annealing is often applied in molecular dynamics calculations (MD; Section 20.1.5) and in the refinement of crystal structures (Section 24.4.6).

20.1.4 Coarse-grained modeling

To calculate the energies for very large systems of several hundred thousand atoms or to increase the time window for MD simulations using QM methods (Section 20.1.5), the number of parameters or degrees of freedom has to be reduced. This is the concept behind coarse-grained methods, which group atoms as *pseudo-atoms* or *beads* that move relative to other beads during the calculations If relative domain movements are studied a bead may represent a whole domain. Large conformational changes may require one bead per secondary structure element. For finer details, suitable groups of atoms are chosen, for example an amino acid residue or a nucleotide. If each amino acid residue or nucleotide is approximated by a bead, a *chain-compatible* bead model results where the distance of the beads is the same as the characteristic distances (Cα-Cα for proteins or phosphate-phosphate for nucleic acids) in the authentic molecule. Higher-precision modeling for proteins is approximated by separate beads for main chain and side chain. Similarly, nucleotides can be broken down into sugar, phosphate, and nucleobase. Phospholipids for membrane modeling can be simplified into beads as sets of 2–4 methylene groups for the alkyl and glycerol chains and one for the polar head group. The representation of solvent follows a similar pattern: several water molecules may be merged into a "super-water", and an ion plus its hydration shell becomes an "ion bead". Beads are assumed to behave similar to atoms as point-like masses obeying Newtonian mechanics. Bonds and angles between connected beads are represented by harmonic springs and harmonic angular potentials. The Lennard-Jones and Coulomb potentials are used to represent long-range interactions. The fewer number of parameters of the bead models not only speed up the calculations, but the energy landscape is also smoother. Therefore, the energetic minimum is calculated in fewer steps. Coarse-grained modeling enables MD calculations on the formation of protein-lipid complexes, diffusion of proteins in membranes, membrane bending, deformation and dynamics of virus capsids, and large assemblies such as the nuclear core complex. Using coarse-grained models, MD simulations on the millisecond time scale can be realized. This is the time range accessible to rapid-mixing kinetic experiments (Chapter 28), and simulations can be validated by comparison with results from these experiments. Techniques that typically deliver medium- to low-resolution data, such as EM (Section 26) or SAXS (Section 23.2.4), often use coarse-grained modeling together with the experimental data (Box 20.1).



20.1.5 Molecular mechanics and dynamics

Energy minimization is just concerned with the end result, the lowest possible energy state, but does neither care about the method nor the path leading to the energy minimum. By contrast, in molecular dynamics (MD), the *trajectory* from one conformational state to the other is of interest, and each conformational state along the trajectory may contain useful biophysical information. MD thus extends the concept of energy minimization to the simulation of molecular motion in a time-dependent manner. Such motions can be anything from simple vibrations (bond stretching and angle bending that give rise to IR spectra; Section 21.4) to the large-scale conformational transitions in protein and RNA folding. Trajectories in MD calculations are described by the time-dependent coordinates, $r_i(t)$, of the atoms i in the system.

In a classical MM approach for MD calculations, Newton's law is used to describe the time-dependent energy function $E(t)$:

$$-\frac{dE(t)}{dr_i} = -\nabla E = F_i(t) = m_i \cdot a_i(t) = m_i \cdot \frac{d^2 r_i(t)}{dt^2} = m_i \cdot \frac{dv_i(t)}{dt} \qquad \text{eq. 20.9}$$

We have seen in Section 20.1.2 that the force acting on an atom i is numerically calculated as the derivative, *i.e.* from the change in energy between its position r_i and its position a small distance away, r_i+dr_i. If we know the masses m_i of the atoms, we can calculate their accelerations using Newton's law $F = m_i \cdot a_i$. The acceleration is linked both to the velocity v_i of the atom and to its position r_i: it is the second time derivative of the position of the atom and the first time derivative of the velocity (eq. 20.9). We can therefore calculate both the velocity and the position of the atoms from their acceleration by single and double integration over time. From the definition of the acceleration as the first time derivative of the velocity we start with

$$a_i(t) = \frac{dv_i(t)}{dt} \qquad \text{eq. 20.10}$$

Integration of eq. 20.10 over time yields:

$$\int_{t=0}^{t} a_i(t)dt = \int_{t=0}^{t} \frac{dv_i(t)}{dt}dt \qquad \text{eq. 20.11}$$

$$a_i(t) \cdot t - a_{i0} = v_i(t) - v_{i0} \qquad \text{eq. 20.12}$$

where v_{i0} is the velocity at time $t = 0$ and a_{i0} is the initial acceleration, which is zero at the start of the calculation. Rearrangement of eq. 20.12 yields eq. 20.13:

$$v_i(t) = a_i(t) \cdot t + v_{i0} \qquad \text{eq. 20.13}$$

Similarly, from the definition of the velocity as the first time derivative of the position we have

$$v_i(t) = \frac{dr_i(t)}{dt} \qquad \text{eq. 20.14}$$

and integration of eq. 20.14 over time results in

$$\int_{t=0}^{t} v_i(t)dt = \int_{t=0}^{t} \frac{dr_i(t)}{dt}dt \qquad \text{eq. 20.15}$$



$$\int_{t=0}^{t} v_i(t)\,dt = r_i(t) - r_{i0} \qquad \text{eq. 20.16}$$

Rearrangement of eq. 20.16 gives an expression for the trajectory $r_i(t)$ as a function of velocity $v_i(t)$:

$$r_i(t) = r_{i0} + \int_{t=0}^{t} v_i(t)\,dt \qquad \text{eq. 20.17}$$

Substitution of eq. 20.13 into eq. 20.17 replaces the velocity $v_i(t)$ with the acceleration $a_i(t)$:

$$r_i(t) = r_{i0} + \int_{t=0}^{t} \left[a_i(t)\cdot t + v_{i0}\right] dt \qquad \text{eq. 20.18}$$

Integration of eq. 20.18 yields

$$r_i(t) = \frac{1}{2} a_i(t)\cdot t^2 + v_{i0}t + r_{i0} \qquad \text{eq. 20.19}$$

Eq. 20.19 shows that the positions $r_i(t)$ of the atoms can be calculated from the known set of starting positions r_{i0}, the known set of starting velocities v_{i0}, and the accelerations $a_i(t)$ derived from the differential of the energy function (eq. 20.9). Before briefly discussing some of the actual integration algorithms used in MD, we will take a look at the necessary boundary conditions of an MD simulation.

20.1.5.1 Boundary conditions and solvation

To prepare an MD simulation, all atoms of the starting structure are assigned initial velocities. This effectively adds a kinetic energy term to the potential energy function: the overall energy of the system becomes the sum of potential and kinetic energies. The velocities of the atoms, and hence the kinetic energy of the system, are determined by the temperature (Section 15.3).

$$E_{kin} = \frac{3}{2} RT = \frac{1}{2} \cdot \sum_i m_i v_i^2 \qquad \text{eq. 20.20}$$

The kinetic energy is applied by heating the system to the simulation temperature, whereupon the atoms start to oscillate about their equilibrium positions. The system is then thermodynamically isolated to conserve the total energy, and allowed to evolve. During this process, different atomic positions are calculated by solving the equations of motion. The structures obtained correspond to different energy states, and the movement of the atoms corresponds to different forces, which are calculated in time increments δt on the order of 1–2 femtoseconds (10^{-15} s). The volume and energy of the system and also the number of atoms in that volume are kept constant using boundary conditions (Figure 20.1): no atoms may leave or enter the volume or else the system is not isolated any more. Isolation of the *in silico* system in a thermodynamic sense, *i.e.* energy and mass conservation (Section 1.1), is important but difficult to achieve for hydrated macromolecules because water may attach or detach from the macromolecule and diffuse out of the MD volume. Inclusion of water is crucial for realistic MD simulations because the large dielectric constant of water (ε_r = 80) has a huge damping effect on electrostatic interactions (Section 17.2.2), and thus on the total energy. To prevent water molecules from diffusing out of the MD volume, *periodic boundary conditions* are applied. The macromolecule is placed in a water-filled cubic box, which is then copied in all three dimensions. The coordinates of the atoms in the central box evolve according to the Newtonian equations and are related to all other boxes by simple translations. Each water molecule leaving the central box at one side is replaced by an exact copy entering the box from the opposite side. The net effect is that the water experiences the same interactions as in bulk solvent (it may diffuse freely), but the number of water molecules per box stays the same (Figure 20.1).



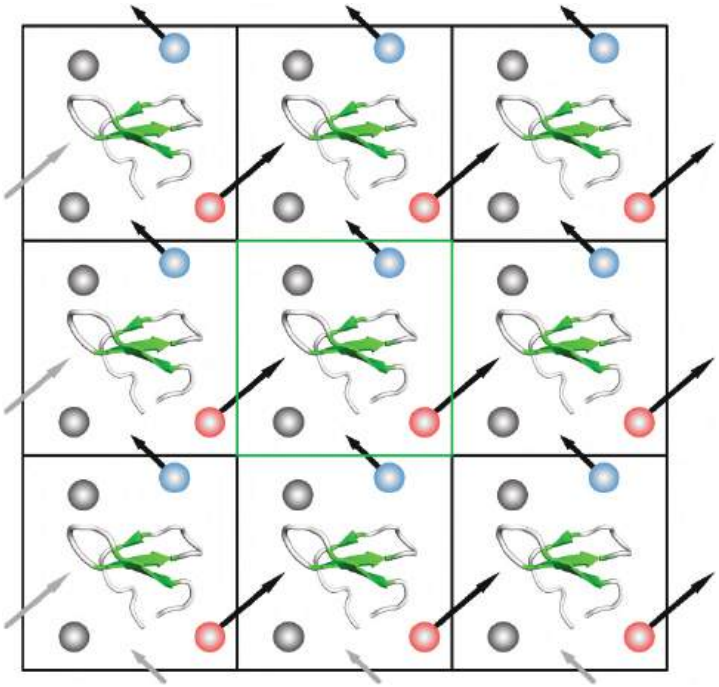


FIGURE 20.1: Periodic boundary conditions for explicit solvent molecules. The protein of interest is at the center of the box, surrounded by solvent molecules (spheres). The number of atoms in the green box stays constant because as soon as one solvent molecule leaves the central box, it is replaced by an identical molecule entering from a neighboring box on the opposite side.

Hydration of a macromolecule can also be implemented implicitly, without specifying each and every solvent molecule. *Implicit solvent models* include extra energy terms describing *hydration* either in the potential energy function or by using a very large number of water molecules and keeping them in place by a *cavity potential function*. Implicit solvent models are computationally less demanding but describe the solvent as an overall environment of the molecule. By contrast, periodic boundaries with explicit solvent, despite being computationally more costly, involve no extra adjustments of the force field and include solvent molecules as part of the MD calculation.

20.1.5.2 Integration of the Newtonian equations

Integration of the Newtonian equations of motion cannot be performed analytically but, as in energy minimization, has to be performed numerically. The integration is broken down into a series of small time intervals δt. At each time point t, $t+\delta t$, $t+2\delta t$, *etc.*, the force is calculated, which yields the acceleration, from which then the velocity and the new coordinates at $t+\delta t$ are derived. The force is assumed to be constant during each time step, but it may change throughout the simulation due to changes in the distribution of potential energy and kinetic energy among the atoms. The integration of the Newtonian equations uses different algorithms, among them the *Verlet integrator*, the *leap-frog* method, and the *velocity Verlet*. Some of these algorithms approximate the Newtonian equations of motion as Taylor series:

$$\boldsymbol{r}(t+\delta t)=\boldsymbol{r}(t)+\boldsymbol{r}'(t)\cdot\delta t+\frac{\boldsymbol{r}''(t)}{2!}\cdot\delta t^2+\cdots=\boldsymbol{r}(t)+\boldsymbol{v}(t)\cdot\delta t+\frac{\boldsymbol{a}(t)}{2}\cdot\delta t^2+\cdots \qquad \text{eq. 20.21}$$

$$\boldsymbol{v}(t+\delta t)=\boldsymbol{v}(t)+\boldsymbol{a}(t)\cdot\delta t+\frac{\boldsymbol{a}'(t)}{2}\cdot\delta t^2+\cdots \qquad \text{eq. 20.22}$$

$$\boldsymbol{a}(t+\delta t)=\boldsymbol{a}(t)+\boldsymbol{a}'(t)\cdot\delta t \qquad \text{eq. 20.23}$$



where $\boldsymbol{r}(t)$, $\boldsymbol{v}(t)$, $\boldsymbol{a}(t)$, and $\boldsymbol{a}'(t)$ are the coordinates, velocity, acceleration, and the first derivative of the acceleration, respectively, for every atom at the time t. As in energy minimization, the Taylor series expansion is often truncated at the quadratic term. The popular Verlet integrator for example uses the following Taylor series approximations for the atomic positions at $t \pm \delta t$:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\cdot\delta t + \frac{\boldsymbol{a}(t)}{2}\cdot\delta t^2 \qquad \text{eq. 20.24}$$

$$\boldsymbol{r}(t-\delta t) = \boldsymbol{r}(t) - \boldsymbol{v}(t)\cdot\delta t + \frac{\boldsymbol{a}(t)}{2}\cdot\delta t^2 \qquad \text{eq. 20.25}$$

Addition of eq. 20.24 and eq. 20.25 eliminates the velocities and yields an equation describing the new atomic positions at time $t+\delta t$:

$$\boldsymbol{r}(t+\delta t) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t-\delta t) + \boldsymbol{a}(t)\cdot\delta t^2 \qquad \text{eq. 20.26}$$

Thus, the Verlet integrator uses the positions and accelerations at time point t together with the positions at time point $t-\delta t$ from the previous step to calculate the new positions at time point $t+\delta t$. Velocities are not explicitly used, but can be calculated by subtraction of eq. 20.24 and eq. 20.25:

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t+\delta t) - \boldsymbol{r}(t-\delta t)}{2\delta t} \qquad \text{eq. 20.27}$$

By contrast, the *leap-frog* algorithm does explicitly include the velocities in the calculations and assumes the following relations:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}\left(t+\frac{\delta t}{2}\right)\cdot\delta t \qquad \text{eq. 20.28}$$

and

$$\boldsymbol{v}\left(t+\frac{\delta t}{2}\right) = \boldsymbol{v}\left(t-\frac{\delta t}{2}\right) + \boldsymbol{a}(t)\cdot\delta t \qquad \text{eq. 20.29}$$

The acceleration is again obtained from the force, then the velocity is calculated at the intermediate time step $t+\delta t/2$ using the velocity from the previous time step $t-\delta t/2$ to reduce errors introduced by the assumptions made for the velocity in the first step. Finally, the atomic positions are updated by modifying the previous positions $\boldsymbol{r}(t)$ using the intermediate velocity and the next iteration can be started. The name leap-frog stems from the ½ time step intervals at which velocity and position alternate. A disadvantage of the method is that the total energy of the system at a given time cannot be calculated because position and velocity are not synchronized.

Finally, the *velocity Verlet* algorithm returns $\boldsymbol{r}(t)$, $\boldsymbol{v}(t)$, and $\boldsymbol{a}(t)$ at the same time with high precision, but at the price of longer computation time, using the following relationships:

$$\boldsymbol{r}(t+\delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\cdot\delta t + \frac{1}{2}\boldsymbol{a}(t) \qquad \text{eq. 20.30}$$

$$\boldsymbol{v}(t+\delta t) = \boldsymbol{v}(t) + \frac{1}{2}\left[\boldsymbol{a}(t) + \boldsymbol{a}(t+\delta t)\right]\cdot\delta t \qquad \text{eq. 20.31}$$

Whichever integration method is used, the stability of the coordinates as a function of time (the trajectory) is gauged against a reference trajectory calculated from the Gibbs free energy. The



reference trajectory strictly conserves energy as is required for any MD calculation performed at constant temperature. If the two trajectories diverge, this is an indication of failure to conserve kinetic and/or potential energy during the simulation and the results will not be useful. The stability of the integration process is improved by using small time increments δt for the integration, typically 1–2 fs for protein simulations.

20.1.5.3 Trajectory analysis

The trajectory describing the motion of the molecular system is the series of conformations calculated over the total time of the simulation, including the coordinates, the velocities, and the accelerations of the atoms, from which energies and forces may be calculated. If the coordinates change substantially during the simulation, this is an indication that the starting structure was of high energy. This is the case for *de novo* protein folding and NMR structure determinations, which start from an unfolded polypeptide chain and end at a folded, compact structure. Sizeable changes in atom position are also observed for crystal structures that contain errors or for proteins that have several (alternative) low-energy conformations. In these cases, energy minimization of the starting structure is essential prior to starting the MD simulation. The trajectory is also analyzed in terms of time and ensemble average structures, which are compared to the starting structure. Time and ensemble averages are an important concept not only in MD but also for example in single molecule fluorescence microscopy (Section 25.7.3). A time average is the mean of a variable over a certain time span. The time average is often generated using data from a single experiment. An ensemble average, on the other hand, is generated from independent experiments, *e.g.* independent MD trajectories or independently calculated structures from NMR, SAXS, or FRET restraints. In MD, either type of averages allows extraction of information from the microscopic energy states of the atoms in the system to describe the energy of the macromolecule. This requires the MD simulation to run long enough (many picoseconds), and at each time point $t+\delta t$ the system must be allowed to reach equilibrium. The resulting trajectory then contains all possible positions and velocities accessible to the system, *i.e.* the molecule samples the entire conformational space. Under these conditions, the *ergodic theorem* applies, stating that an ensemble average at a specific time point is equivalent to a time average over the "infinitely long" trajectory. The theorem also applies *vice versa*: a sufficiently large number of measurements will have a time average that is a good approximation to the equilibrium ensemble average. Importantly, the energies calculated from time and ensemble average structures can be compared to experimentally determined energies, *e.g.* from isothermal titration calorimetry (Section 30.1), providing a link between theory and experiment.

The end point of an MD simulation is not fixed but only limited by the available computational resources. Time frames from nanoseconds up to a few microseconds using time steps in the femtosecond range are achievable using several days to weeks of computation time, often aided by parallelization. These trajectories are sufficient to describe the dynamics of small proteins and of double-stranded DNA but insufficient for large macromolecular assemblies or slow conformational changes in *e.g.* protein folding (Table 18.12) or the tracking of diffusing molecules (Section 25.7.3). To access the slower biologically interesting processes, the MD calculations need speeding up by simplifying either the force field or the system itself. Since the most time-consuming part of a MD simulation is calculation of the non-bonded energy terms, a cut-off distance can be introduced into the function to disregard all interactions between two atoms sufficiently separated in space. Experimental data can be included in the potential energy function to serve as powerful restraints. The number of variables in the system itself may be drastically reduced by coarse-grained modeling (Section 20.1.4). While these adaptations make the simulations on the microsecond scale computationally feasible, slow conformational transitions in proteins such as allosteric changes, oligomer assembly, and protein folding currently still remain difficult for all-atom MD simulations.



20.2 PRINCIPLES OF NEURAL NETWORKS AND MACHINE LEARNING

In this Section, a very brief introduction on machine learning is given. Machine learning algorithms advance biophysics in several fields, including molecular dynamics and protein modeling. The term *artificial intelligence* (AI) describes tasks performed by machines that would normally require a human brain. Examples of such tasks are speech recognition, language translation, and *computer vision*. Computer vision is a collective term for all sorts of image analysis for patterns that usually elude the human eye, including X-ray diffraction images, cryo-EM micrographs, and images from fluorescence microscopy. *Machine learning* is a field within AI devoted to developing methods that use prior data, termed the *training set*, to improve their performance on unknown data in these tasks. Many machine learning algorithms use statistics gained from the training set to build a model that continues to evolve during the training. Machine learning algorithms are thus more versatile than a program made for an explicit task. Learning procedures can involve multiple layers of abstraction that progressively extract higher-level information from raw input data. Many layers are a hallmark of *deep learning*. An example could be image processing in cryo-EM where an early layer identifies edges of objects and a subsequent layer groups the objects into classes. Deep learning calculations are done in *artificial neural networks*, often just called *neural networks*. A biological neural network is a population of neurons connected by synapses, in which the neurons perform some type of calculation and pass the result on to nearby neurons. By analogy, an artificial neural network is a set of interconnected units or *nodes* that perform the calculation and pass the result on to nearby nodes. Nodes are typically organized in layers, and several layers make up a *module* performing the different tasks of deep learning. Layered neural networks are typically organized into input layer, hidden layers, and output layer. Depending on how the layers are connected, different networks suitable for different tasks result (Figure 20.2). In a fully connected neural network, all nodes of the previous layer are connected to all nodes of the next layer. In a *feed-forward network*, only a subset of the nodes is connected. As a result, information transfer becomes directional, it cannot move backwards. Finally, in case of "pooling" of nodes, a group of nodes in the current layer connects to a single node in the following layer, thus continuously reducing the number of nodes across layers. Deep neural networks can contain hundreds of layers, so the type of network determines the computational requirements.

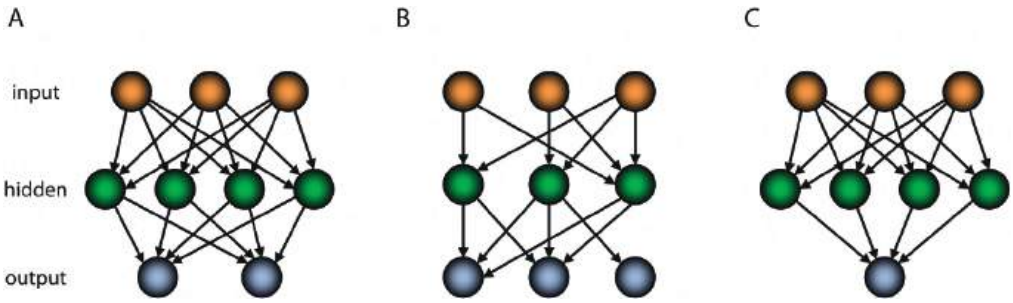


FIGURE 20.2: Principle of a neural network. A: Fully connected neural networks have all nodes from one layer connected to all nodes of the following layer. B: A *feed-forward network* has lower dimensionality because only subsets of nodes from one layer connect to subsets of nodes in the next layer. C: Networks with pooling reduce the complexity and are thus computationally efficient.

Neural networks are suitable for pattern recognition tasks in sequential data, which include quite a diverse set such as protein and nucleic acid sequences, speech, hand writing, and gestures. Networks for pattern recognition, which are typical for biophysical applications, are trained by *supervised learning*. Supervised learning requires training data where the input and output are known. Such data are often called "labeled data". For example, a certain piece of raw data (a protein sequence maybe) enters the input layer, is processed in the hidden layers, and a result is produced by the output layer (a structural model in this case). The aim of the training is to establish a general model for the network: an association between input and output, weighted by



probabilities. The output result is compared to the known true result of the training example. The difference between prediction and true result (known from the training dataset) is the error. The error is fed back iteratively into the neural network for more rounds of calculation. Exactly how the error affects the learning process depends on the *learning rule* for the training. Several learning rules exist for training neural networks that differ in how the error affects the weighting schemes between nodes. Two examples of learning rule for the supervised training of neural networks are the "correlation learning rule" and the "delta learning rule". The correlation learning rule states that if two neighboring nodes produce a similar result, the weight between these nodes should be more positive for the next round of prediction, and *vice versa*. In delta learning, the error of the prediction is minimized by a steepest gradient algorithm over all the patterns in the training data. Once a learning rule has been pre-defined for the training, the probability weights for the input/output association of nodes for the next cycle are adjusted by that rule with the help of the error value from the previous cycle. A function related to the error, known as the *cost function*, is minimized, similar to the minimization methods we have discussed in Section 20.1.2. For the delta learning rule the steepest gradient algorithm is used to minimize the cost function. For the case of predicting structural models, the cost function could be the rmsd (Section 31.12) between model and true structure. Once the error remains minimal, even if additional input training data is provided, the training might be considered completed. However, there are two scenarios that may impair the training of the network, the vanishing gradient and the exploding gradient problems. The gradient is a vector, calculated as the partial derivatives of the loss function with respect to the parameters of the model being trained. The larger the gradient, the more strongly will the loss function increase, so the aim of the training is a small gradient. The direction of the gradient shows which weight change influences the loss function strongest. Using the gradients, the weights are updated and back-propagated through the network. Back-propagation means that weights are updated first for the final layers, then for the intermediate, and so on until the first layers. During back-propagation, gradients are successively multiplied through the layers, with the consequence that for small gradients, the products of the gradients tend to become close to zero (they "vanish"). As a result of tiny gradients, the weights for the first layers in the network won't change at all, locking the network and prohibiting it from learning. The vanishing gradient scenario is therefore of particular concern in networks with many layers where many multiplications of gradients happen. The opposite is true for large gradients, which become ever larger by multiplication (they "explode"). Huge gradients lead to excessive changes in weights during back-propagation, with concomitant numerical instability during training. Both problems, vanishing gradient and exploding gradient, may be mitigated or avoided by choosing the best learning rule for the model. Another solution to the learning impasse is using a module containing an *attention mechanism*. The computer equivalent of human cognitive attention is the ability of an algorithm to concentrate on a discrete aspect of the raw input data. Such modules are also known as *transformers*, and they are at the heart of language interpretation such as ChatGPT (Generative Pre-training Transformer). The result of the transformer is, on the one hand, an enhanced weighting of some small but useful parts of the input data, and, on the other hand, a downweighting of a large part of the data that do not contain useful information. The attention mechanism in transformers achieves this by encoding the raw input data such as language or sequences as vectors. The vectors then describe the data as "word in a sentence" or "position of a residue in a sequence". The vectors are grouped into three sets called queries, keys, and values, where the query describes the property to find in the data (*e.g.* spatial proximity of residues in a structure), the key defines some sort of attribute the data has (*e.g.* interaction), and the value is the information associated with the data (*e.g.* favorable or unfavorable). A dot product (Section 31.9.1) between query and key vectors is calculated, which results in a single number that is large if the vectors are parallel and zero if they are perpendicular to each other. The numbers are normalized and scaled using a *softmax* function (eq. 20.32).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}} \qquad \text{eq. 20.32}$$



The exponential function is applied to each of the n data values z_i and divided by the sum of all the exponentials. This normalization ensures that the sum of all $\sigma(z)_i$ is unity so the converted set of numbers can be interpreted as probabilities. The feedback and re-assessment of probabilities during the supervised learning process results in a model that can now extract information from a large dataset of a similar type it was not trained on. Advantages of networks trained by supervised learning are a clear objective combined with high accuracy and interpretability of the predictions.

These properties distinguish machine learning from "just sophisticated statistics". Still, as the training set may be biased, there is always the danger that the output of a machine trained by supervised learning is also biased. In addition, if the raw input data is unusual (under-represented in the training set or even absent from the training set), the output may be inaccurate or just plain wrong. One of the most successful applications of machine learning in structural biophysics was the three-dimensional prediction of protein structures from sequence, which we turn to next.

20.3 SIMULATING PROTEIN FOLDING

The Anfinsen experiments on RNase showed that the three-dimensional structure is an inherent property of the protein sequence at given environmental conditions (Section 18.5.2). An early dream in protein science has been to calculate the structure of a protein from its sequence alone, to decipher the "protein code" in analogy to the genetic code. Historically, the best approaches to convert a sequence into a model needed a known template structure of a related sequence. Recent machine learning algorithms such as AlphaFold2 or RoseTTAFold do away with the requirement of a template structure and can calculate *ab initio* protein models from sequences alone (Section 20.3.2). Structural information, if available, is incorporated in the prediction but not necessary. We will discuss template-based approaches first, before turning to *ab initio* methods.

20.3.1 Homology modeling

Until 2020, the most successful approach to predict structures from sequences was homology modeling. In this method, a target sequence is aligned with sequences of close homologs. For at least one of the sequences a high-resolution (crystal) structure must be available. This *parent structure,* or superimposed group of parent structures, is then used as a template for the new structure. If there are several related structures, a conserved core can be extracted that is more reliable to build upon than a single structure. Those residues that align well between the sequences are likely modeled with high confidence, sometimes on the accuracy level of side-chain rotamers. The side chains of all other residues are placed in energetically favorable conformations. Any clashes between the side chains are then removed by energy minimization (Section 20.1.2). For query sequences with insertions compared to the template, the last step is the modeling of surface loops. Very short loops often follow the conformations of canonical turns (Section 18.3.3). Longer loops of eight and more residues may be modeled either by template-based techniques, or from loop databases, or by *ab initio* modeling. Suitable template loops in this *spare parts approach* have similar length, similar connections to the protein, or similar sequence. Loop databases, such as for the complementarity-determining regions of antibodies, are used to look up the possible conformations of whole loops or smaller parts thereof, which are then grafted onto the core structure. In *de novo* modeling, the loops may be grown residue by residue from their anchor points on the core, until a closure of the loop can be found. An alternative route is to create a loop by adding dipeptides of sensible stereochemistry followed by refinement, or to deconstruct a larger loop until it fits the length, sequence, and anchor points of the core. In all approaches, Ramachandran restraints (Figure 18.21) are useful to ensure a sensible backbone conformation of the loop modeled. Homology models for which a crystal structure became available later showed that the accuracy can vary substantially, from ca. 0.1 nm rmsd for ultra-short loops of three residues to about 0.4 nm rmsd for loops of eight or more residues. Homology modeling is particularly successful for antigen-binding fragments (Fabs; Figure 24.5) where thousands of crystal structures exist: the core structure of the immunoglobulin domains is very similar and the myriad



of loop conformations that the complementarity-determining regions adopt constitutes a large loop database. For other folds with much fewer structural data available, other algorithms were developed and their efficiency was compared in a competition called CASP (Box 20.2).

BOX 20.2: CASP (CRITICAL ASSESSMENT OF STRUCTURE PREDICTION; PREDICTIONCENTER.ORG).

The CASP challenge is a biennial competition since 1994 where dozens of research groups have their modeling algorithms compete for the best three-dimensional structure predictions starting from a protein sequence. The true structure of the protein is known but withheld until after the competition and used as a benchmark for how accurate the predictions are. Traditional disciplines in CASP are template-based and template-free methods with the latter being more challenging since no structural information at all is available to guide the predictions. Unsurprisingly, the success rate of template-free algorithms is smaller than for predictions that can make use of a template. However, the template-free approach is ultimately the more important one because not all sequences can be matched to a closely related structure that can serve as a template. Up until the 12th CASP in 2016, the best modeling algorithms predicted about 35–40% of the residues satisfactorily in the template-free discipline. The percentage increased to about 60% in CASP 13 (2018) when a new contestant, Deepmind, introduced the AlphaFold algorithm in the competition. It was the first neural network algorithm employed for CASP. In 2020, a revised version, AlphaFold2, outcompeted every other algorithm by predicting 87% of all residues satisfactorily (Figure 20.3). The average rmsd error of the predictions was 160 pm, which is sufficiently precise and accurate for the models to serve as search models for molecular replacement phasing of diffraction data (Section 24.4.4.3).

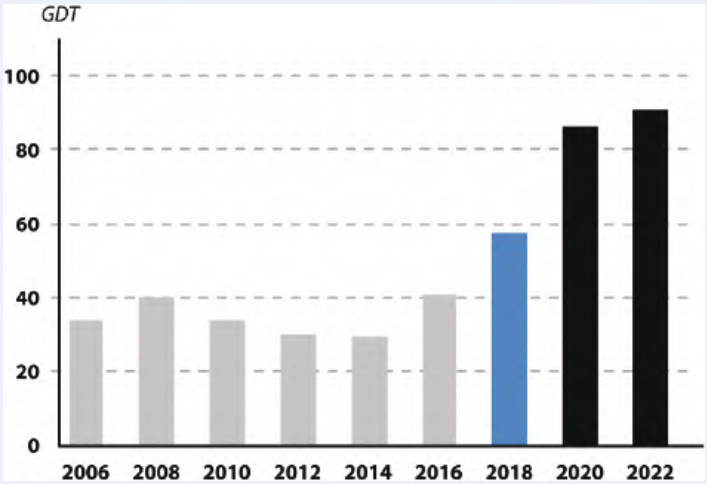


FIGURE 20.3: Success rate for the best modeling algorithms in CASP. Shown is the median accuracy of the predicted models for the best group in the template-free modeling category from 2006 to 2020. GDT is the global distance test, measuring the similarity in percent between the predicted model and a reference structure, in this case the withheld crystal structure (see Section 20.3.3). The first version of AlphaFold entered the competition in 2018 and immediately superseded all other algorithms (blue bar). In 2020, the revised algorithm AlphaFold2 scored 87% GDT in the template-free modeling category, about twice as high as the second best group. Models with scores approaching 90% GDT are close to experimentally determined structures.

(*Continued*)



In 2022, AlphaFold2 did not participate any more, but similar algorithms were developed in the meantime that continued to improve the success rate of the predictions, among them RoseTTAFold, D-I-TASSER, and PEZYFoldings. Other disciplines that are even more difficult than modeling of a single polypeptide sequence have been included in CASP. These include modeling of multimeric complexes, ligand-binding predictions, inter-domain interaction predictions, and RNA folding. Modeling of multimeric complexes was already part of the CASP challenge since 2016, but large progress was made only in 2022.

20.3.2 Sequence-based structure prediction

During several decades of structure prediction algorithm development, progress was limited to template-based methods. Apparently, the information contained in an experimental structure related to a target sequence was essential to obtain a useful model for that sequence. This is not necessarily surprising because protein structure is much more conserved than sequence. As of 2022, the UniRef90 database had > 144 million total sequences, where "90" signifies that only sequences are considered that are less than 90% identical to other sequences. By contrast, only about 72000 structures in the PDB are different at this 90% level. It follows that sequence space is about 2000-fold larger than structure space: an individual structure contains much more information than a sequence because many different residue pairs can lead to similar three-dimensional arrangements. In recent years, novel neural network algorithms have been developed that do not need any structural information to convert a sequence into an accurate molecular model. Instead, they use co-evolution information of residues from multiple sequence alignments (MSAs). As it turns out, co-evolving residue pairs alone are sufficient to guide modeling of sequences into structures. The underlying hypothesis connecting MSAs and structure is that residue pairs that are close in structure space should co-evolve in sequence space. The connection between co-evolution and spatial proximity is intuitive: if two residues are close in space and interact with each other, natural mutation of one residue to a side chain with resulting unfavorable interactions will generate selective pressure for the other residue to mutate as well so that the interactions of the mutated residue pair are favorable again. Taken the other way around, if co-evolving residue pairs can be identified from MSAs, it should be possible to calculate a structural model for the sequence just by using the residue pairs as proximity restraints (Figure 20.4).

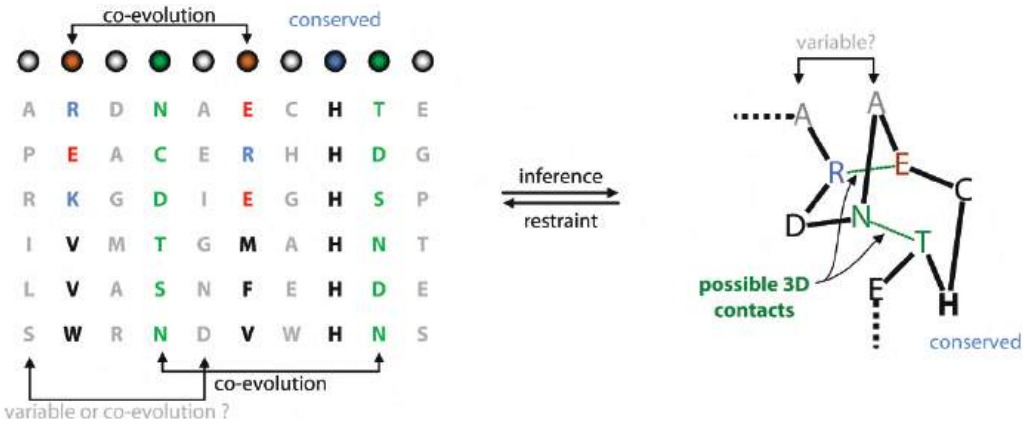


FIGURE 20.4: Spatial proximity of co-evolving residue pairs in multiple sequence alignments. On the left, an MSA is shown based on a conserved histidine residue. Based on this alignment some columns with matching side-chain properties such as hydrophobic/hydrophobic or positive/negative charges indicate a possible co-evolution while for other pairs of columns there is no clear indication of co-evolution. These correlations are used as restraints for generating a first structural model (right-hand side). If the first model is accurate enough, it indicates more residue pairs (*i.e.* columns in the MSA) that should co-evolve, and these may be used as restraints in the next round of prediction to obtain a better model, *etc*. Individual sequences in the MSA may need shifting or insertion of gaps to incorporate new co-evolving residue pairs suggested by the model. The left and right parts of this figure correspond to the "evoformer" and "structure module" in Figure 20.5, respectively.

The basic strategy of how current sequence folding algorithms work is rather similar. In a nutshell, the input data are sequences and sometimes experimental structures. Co-evolution information from sequences is extracted by a module termed the "MSA transformer". Structural information, if available, is analyzed for residue pairs close in space by a module known as the "pair transformer". Both sources of information are combined in a third module, the "evoformer", and used to iteratively improve the MSA during several mini-cycles. Then the proximity data is passed on to a "structure module" that calculates a first crude three-dimensional model. This first model is analyzed for additional proximity information and the MSA and pair representation are updated iteratively in macro-cycles (Figure 20.5). All of the three modules MSA transformer, pair transformer, and evoformer use attention mechanisms to more strongly weigh the important parts of their input data (Section 20.2).

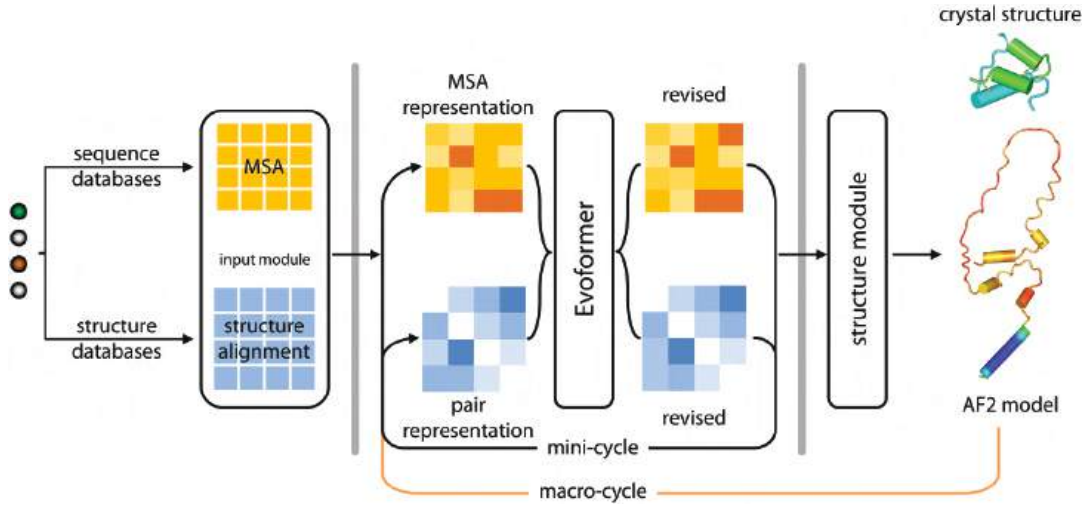


FIGURE 20.5: Outline of the AlphaFold2 algorithm. In this two-tracked neural network, the input modules search the query sequence (spheres on the left) against sequence and structure databases. The Evoformer, a combination of MSA transformer and pair transformer modules known from language modeling, uses both MSA and residue pairs from template structures to build a relationship between residues, then revises the MSA and residue pair representations iteratively until they contain maximal information on thousands of residue relationships (mini-cycles). The structure module uses these relationships to produce a first structural model, from which additional interacting residue pairs are identified. This information is cycled back into the Evoformer for a total of three macro-cycles to arrive at the final model. Most but not all of the models are predicted with high confidence. Insulin (UniProt ID P01308) is an illustrative case for which the prediction is wrong: the model is colored according to the pLDDT value, a measure of local confidence (Figure 20.6) where blue is high and yellow/red is low confidence. The correct crystal structure is shown on top.

MSAs of related sequences are ideally very large, up to thousands of sequences, and while computationally demanding, large MSAs make the prediction of co-evolving residue pairs quite robust. By contrast, the minimum number of sequences required for an MSA to reliably identify co-evolving residues is about 30. In such a limiting case, the MSA may not be diverse enough to calculate an accurate model without the help of prior experimental structures. Only if the MSA contains at least a few hundred sequences can the prediction proceed entirely without experimental structural data.

Computationally, an initial MSA is converted into a machine-readable matrix by the MSA transformer. This program takes an MSA as input and assigns to each residue in a column a value of relative conservation. The MSA transformer then identifies correlations between columns of the MSA, establishing evolutionary correlations between residue positions. These correlations are converted into distances and stored in a *pair representation* matrix where each residue is assigned a probability for coexistence in space with any other residue. The pair representation matrix is handed to the pair transformer. This module takes (two-dimensional) distance information and converts it into a three-dimensional model. Should experimental structural information be available in the



beginning of the calculations, the pair transformer extracts the distance information between residue pairs close in space and adds it to the pair representation matrix prior to calculating the first model. In case no experimental structural data is present, the first model is calculated by the pair transformer solely based on the estimated distances between the residue pairs and the probabilities of their co-evolution from the MSA. The first model is always wrong because the initial MSA inevitably contains mis-alignments. The reason for the mis-alignments comes from the fact that related natural sequences are often of different lengths, corresponding to insertions and deletions in the MSA. Without knowing the size of the insertions any initial MSA contains mis-alignments and the search for co-evolving residue pairs is necessarily incomplete. As a result, the further away co-evolving residue pairs are in the sequence, the more difficult they are to identify in a non-optimized MSA. Many tertiary structure interactions are however established by residues far apart in the sequence. Therefore, both the MSA and the calculated model need optimization by the evoformer, which is done by analyzing the model calculated by the pair transformer for residue pairs close in space that were not identified from the previous MSA. Using these additional residue pairs, the sequences in the MSA are shifted or gaps of varying lengths are introduced. Ideally, the updated MSA shows the additional residue pairs to co-evolve. The new residue pairs are included in an updated pair representation matrix that allows the pair transformer to calculate a better structural model. Ping-ponging between these two sources of information, updated MSA and new residue pair suggestions from an intermediate model, iteratively improves both predictions. The cycle is stopped either after a fixed number of iterations or when successive intermediate models do not differ by more than a given threshold or once the model satisfies the pairwise distance restraints given by the co-evolving residue pairs. At this point, the model has no physical meaning but is just an assembly of "residues" labeled as amino acids, sometimes referred to as a *residue gas*.

The next step is performed by the "structure module", which calculates a physically plausible three-dimensional structure: the residues in the "gas" are linked by peptide bonds following Ramachandran restraints, side chains are installed, rotamers are chosen, and the final model is energy-minimized. While the intermediate models from the pair transformer are generated using the two-dimensional pair representation matrix as restraints, the final model is energy-minimized in three-dimensional space using conventional force fields, which adds an additional source of restraints to generate the best physically plausible model. The pair representation from this model is fed back into the evoformer for another macro-cycle of calculation. Convergence is reached once the calculated energy of the final model falls below a certain threshold. The individual modules of the neural network were well-known before, but the iterative connection between them in the context of machine learning was the key new step in AlphaFold2 that led to an improvement from 40% correctly predicted residues in 2018 to > 90% for related algorithms in 2022.

AlphaFold2 serves as the example algorithm here, but there are several others of similarly high predictive power, including RoseTTAFold, D-I-TASSER, and PEZYFoldings. All of these are ML algorithms trained by supervised learning. For example, one version of AlphaFold2 was trained on protein sequences and structures released before February 15, 2021. Newer structures that have not contributed to the training therefore can be used as benchmarks to assess possible bias in the predicted models. Bias is a known problem in supervised learning. Supervised learning of a neural network is performed using a labeled dataset of known correct results (Section 20.2). This learning method always bears the risk that the output of the trained network resembles the training set too much, hence it may be biased.

A further point of caution relates to the folding pathway for the models. While the "protein folding code" has finally been deciphered computationally, the Levinthal paradox (Section 18.5.3) still exists. The *in silico* folding pathway of the sequence during the calculation is entirely non-physical and does not reveal any possible folding intermediates. This is because the ensemble of residues is not restrained by amide linkages during the first rounds of modeling by the pair transformer. The final model often makes sense, but not its intermediates.



20.3.3 Quality assessment of predicted three-dimensional protein models

Assessment of the quality, or accuracy, of a predicted model is essential to judge its usefulness for generating hypotheses that may be tested by experiments. Several tests are performed routinely, depending on whether a reference structure is available or not, and depending on whether the global or the local accuracy of the prediction needs scoring. The tests outlined here are superior to a simple rmsd calculation and include the global distance test (GDT), the predicted Local Distance Difference Test (pLDDT), and the Predicted Aligned Error (PAE). As the names imply, GDT is a global metric and pLDDT (and also PAE) are local quality metrics.

If a reference structure is available, the predicted model can simply be superimposed on that reference and an overall rmsd value may be calculated. The smaller an rmsd, the better the agreement between model and reference. Rmsd values are insightful for compact, single-domain proteins where most corresponding residue pairs have similar distances in the superposition, *i.e.* the distance distribution between Cα-atoms is rather narrow. They are less useful if the molecules superimposed have large mobile surface loops or consist of several loosely connected domains that may change their relative positions. The global distance test (GDT) is a more fine-tunable measure for the similarity between two structural models than rmsd. GDT measures the percentage of residues in a predicted model that are within a certain threshold distance from their corresponding residues in the reference structure. The two structures being compared are first superimposed roughly, either by matching common sequence stretches, by conserved residue pairs, or by secondary structure elements. Next, the structural alignment is improved iteratively by small adjustments to the parameters used for the superposition until a pre-defined criterium is reached, *e.g.* convergence of the coordinate shift between successive cycles or simply the maximum number of iterations. The result is a superposition with a minimal global distance between model and reference, resulting in the smallest overall rmsd. For GDT calculation, distances between corresponding Cα-atoms that fall below a pre-defined threshold are tallied and put on a percentage scale. CASP (Box 20.2), for example, calculates four GDTs with distance cut-offs 1, 2, 4, and 8 Å. The four GDT scores are averaged into a total score termed GDT_TS. If the superposition is bad, GDT_1 for the 1 Å cut-off will be small while GDT_8 for the 8 Å cut-off will be large, resulting in a small GDT_TS. For a close superposition, all GDT-values will be large, including GDT_TS. The advantage of GDT_TS over rmsd is that GDT_TS is less sensitive to strongly deviating regions (often loops). A disadvantage is its limitation to single domains. Being a global measure for accuracy, GDT cannot account for changes in the relative domain orientation of multi-domain proteins. Variants of the GDT calculations have been developed that are not limited to Cα-atoms but take side chains or even all atoms into account.

In contrast to the global measure GDT, the predicted Local Distance Difference Test (pLDDT) assesses the local "correctness" of a structural model. A sphere of 1.5 nm diameter is drawn around each residue in the model and this local environment is compared to environments expected from experimentally determined structures. pLDDT is the output of a deep learning model that is trained on experimental data, including the relative orientations of side chains, rotamers, packing, and distances. The calculated pLDDT values for each residue are set on a scale from 0% to 100% and placed in the *B*-value column of the model PDB files. But in contrast to *B*-values (Section 24.4.6), higher numbers for pLDDT correspond to a higher positional accuracy (Figure 20.6). Values > 90% stand for very high confidence that the local environment is correct, 70–90% is confident, 50–70% correspond to low confidence, and any residues scored < 50% have very low confidence. An example of low confidence is given for insulin in Figure 20.5. The pLDDT values of all residues may also be aggregated into a single number describing the overall confidence of the model prediction.

Due to the way pLDDT is calculated, it does not involve structure superposition and hence is not a measure for how well the predicted and "true" structures align. Since the true structure is usually unknown in modeling (the CASP challenge is an exception; Box 20.2), pLDDT is superior to rmsd or GDT, which both require a reference structure. In contrast to the standard GDT (Figure 20.3), which is limited to Cα atoms, pLDDT also includes all atoms in the quality assessment.



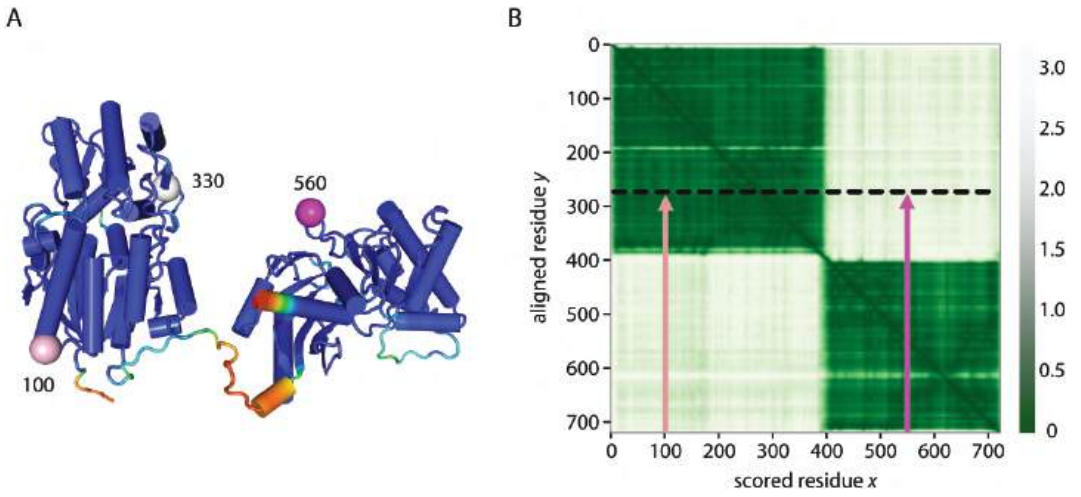


FIGURE 20.6: Confidence matrices for local model quality. (A) pLDDT ranges from 0% to 100% and is usually colored on the structural model according to the confidence it represents for the residue placement. Dark blue is for > 90% (very high confidence), cyan for 70–90% (confident), yellow for 50–70% (low confidence), and orange for < 50% (very low confidence). The structure shown has two confidently modeled domains but the linker between them has low pLDDT. (B) PAEs range from 0 nm to 3.1 nm distance and are shown as a heatmap from dark green at 0 nm error to pale green at an arbitrary cut-off of 3.1 nm with residue numbers on both axes. The diagonal from top left to lower right compares the residue with itself, so it has always a value of 0 nm, corresponding to 100%. The intensity of the green color in position *x,y* is the expected distance error in residue *x*'s position, when both the prediction and the "true" structure are aligned on residue *y*. Large green squares correspond to compact domains. The two domains in this model have no off-diagonal low PAE region, indicating the domains do not interact with each other (see Figure 20.7 for inter-domain interactions). The PAE of residue position $x = 100$ when aligned with position $y = 330$ is small (dark green, tip of pink arrow), so these residues are within the same domain. By contrast, positions 330 and 560 have high PAE values (light green, tip of magenta arrow), which means the residues are on separate domains.

A drawback of pLDDT is its inability to address domain juxtapositions. In multi-domain proteins, each domain can have high pLDDT percentages but the overall rmsd to a "true" structure can still be high due to differences in the relative arrangement of these domains. The example model drawn in Figure 20.6A shows two well-predicted domains connected by a linker that is not well predicted. A high pLDDT for domains in a multi-domain protein therefore does not automatically mean a high confidence for their relative juxtaposition (Figure 20.6).

For evaluating the predictions of inter-domain protein-protein interactions the Predicted Aligned Error (PAE) is a better measure. PAE is equivalent to pLDDT within domains but becomes more instructive when comparing residue pairs from different domains. The PAE reports, in the form of a matrix, the positional error of residue *x* when model and true structure are aligned on residue *y* (Figure 20.6). In other words, the PAE measures the confidence in the relative position of residue pairs *x,y* in the model compared to the "true" structure. Because the true structure is unknown, a surrogate true structure is extrapolated from experimental structures in the PDB: the algorithm calculating PAE was trained on structural data so the superposition between the predicted model and a hypothetical true structure gives similar results as if the true structure were known.

Of note, the PAE plot is not an inter-residue distance map but the intensity of the green color in the heatmap only indicates the expected distance error between the two residues. Values for PAE range between 0 nm and 3.1 nm, with ≤ 0.5 nm indicating faithful placement of the residue pair and 3.1 nm being an arbitrary cut-off for high error estimates. If the residues in the pair are from different domains and PAE is small, it is likely that the two residues interact across the two domains. Especially if whole sequence stretches from different domains show consistently small PAE, PAE becomes a confidence measure for domain packing. Such residue stretches show as off-diagonal areas in the PAE plots and may indicate biologically relevant protein-protein interactions (Figure 20.7).



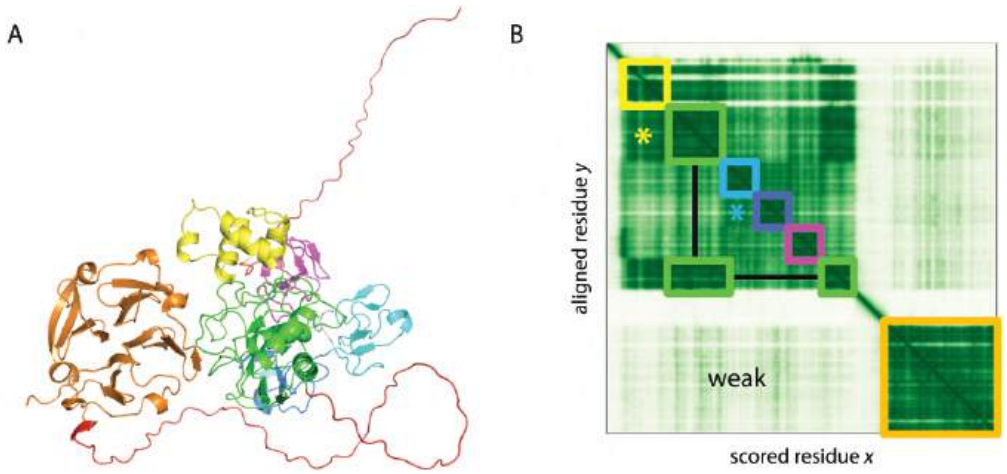


FIGURE 20.7: Protein-protein interactions and disorder prediction. (A) A multi-domain protein is shown with its individual domains colored. The N-terminus (top right) and a large loop (bottom) are predicted as disordered. (B) The corresponding PAE plot with the sequences of the domains in (A) boxed with the same colors. Off-diagonal dark green rectangles indicate predicted inter-domain interactions. For example, the green domain is composed of two separate sequence stretches whose interactions lead to an off-diagonal box. The yellow domain interacts with the green domain (yellow off-diagonal asterisk), and the light and dark blue domains interact as well (light blue asterisk). Only weak interaction, if any, is predicted between the orange domain with the rest of the protein.

20.3.4 Biophysical value and limitations of predicted structural models

There are some unexpected novel features in the models generated by AlphaFold2 and related algorithms. These include disorder prediction, multiple conformational states, and oligomer formation. We will describe these three aspects before turning to the caveats of predicted models.

Although never designed for it, AlphaFold2 and related algorithms turned out to predict protein-protein interactions and protein disorder quite faithfully (Figure 20.7). A pLDDT value < 50% over a stretch of residues predicts disorder in this region. Such regions surprisingly often correspond to the low complexity regions (LCRs) identified from primary sequence analysis (Section 18.2), among them repetitive sequence motifs such as those found in collagen. If most of the sequence or even the entire sequence have low pLDDT, the protein might be intrinsically unfolded. Such proteins may still fold in the presence of other interaction partners, *e.g.* ribosomal proteins that fold only in the presence of rRNA. Smaller sequence stretches of low pLDDT often include loop regions between secondary structure elements and the termini, indicating possible flexibility also in solution. Although never intended for this purpose, regions of small pLDDT are good indicators for disorder and useful in the design of compact protein constructs for biophysical studies.

Many regulatory proteins, including G-proteins, GPCRs, and kinases, have several distinct biologically relevant conformations. Early versions of folding algorithms were implemented to generate a single conformation of a protein model, namely the one with the lowest calculated energy after minimization. It was impossible to predict which of the conformations output, if any, corresponded to a biological role. Current implementations of AlphaFold and RoseTTAFold output several models, which can be validated by comparing their conformations to biological or biophysical data (*i.e.* does the conformation make sense?).

Interaction partners were not present in the training set for the sequence folding algorithms, so modeling of protein-protein complexes was initially not a standard task. Users soon found out that to generate a model for an oligomeric complex, it was often sufficient to submit a single sequence to the calculation where the binding partners were connected by LCRs. In some instances the LCRs impeded accurate folding of the complex, even if the order of the binding partners and the lengths of the connecting LCRs was changed. Recent implementations of AlphaFold and RoseTTAFold now have dedicated tasks for multimer folding. Currently, predicting structures of hetero-oligomers accurately is more difficult than the prediction of homo-oligomer structures.



There are also a number of limitations to be aware of when using ML-generated protein models. These are placement of flexibly linked domains, binding of non-protein components, post-translational modifications, and the effect of natural mutations. Unfortunately, all of these are of considerable biological interest, and development of novel algorithms seeks to answer biological questions that cannot currently be addressed by the predictions.

For example, if the domains in a protein are connected by flexible linkers and the algorithm does not predict inter-domain interactions (Figure 20.6), placement of the domains will not necessarily be at biologically sensible positions. In membrane protein predictions an intra- or extracellular domain may occupy space that is occupied by membrane *in vivo*. Membranes are neither part of the training set nor can their location be specified for the predictions yet.

The same is true for all non-proteinaceous components: ligands, metal ions, and post-see above modifications are absent from the predicted models because the training set did not include them either. Models calculated from sequences without taking ligands or post-translational modifications into account therefore tend to show bias toward the (incomplete) training set. A good example is insulin (Figure 20.5), which requires proteolytic processing and formation of disulfide bonds to fold. Mature insulin consists of two polypeptide chains generated by proteolysis from a precursor sequence. The two chains are connected by three disulfide bonds and insulin can also form a hexamer. Folding the sequence of the un-processed insulin failed because while there are many insulin crystal structures available in the PDB, none is present of the precursor insulin sequence and it is not sure whether the precursor folds into a defined structure at all.

Maybe somewhat surprisingly, predicted models for proteins binding to co-factors and metal ions have very similar conformations for the side chains that normally coordinate these small entities (Figure 20.8). Again, this might be an expression of bias coming from the structural part of the training set, which contains the apo-protein structures stripped of their ligands. Another algorithm, AlphaFill, was developed to place ligands back into the AlphaFold2 models, and yet another algorithm, DragonFold, promises joint co-folding of protein sequences and ligands into a protein-ligand complex.

Single point mutations are also a case where the model predictions are poor. If the hydrophobic core of a protein is "mutated" *in silico* and the sequence passed through the calculation, the astonishing result is a model that appears overall reasonable, despite being physically impossible (Figure 20.8). Likewise, the effect of natural point mutations cannot yet be reliably predicted by folding of the mutated sequence. Since the success of a prediction hinges on a large MSA, a single point mutation will not strongly affect the alignment, and hence will not change the final model much.

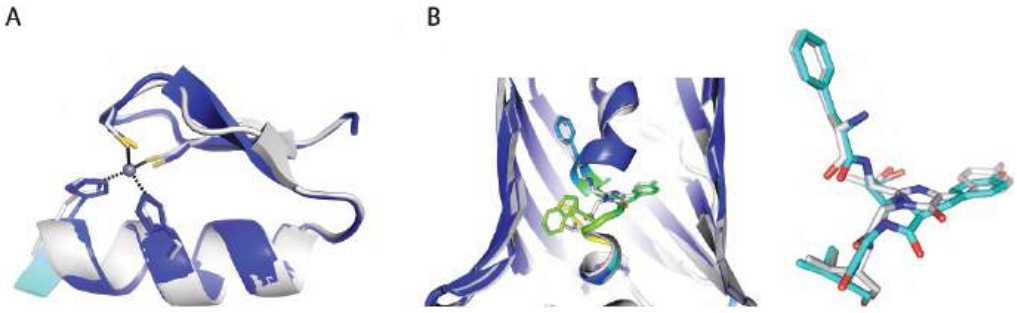


FIGURE 20.8: Comparison of crystal structures and AlphaFold2 models. Crystal structures are depicted in gray, and models in blue according to their pLDDT. (A) ZIF268 C2H2 zinc finger from Figure 18.12A. The training set for sure included zinc finger structures but without the metal ion, yet the side chains of the model (blue) are very close to their positions in the crystal structure. The rmsd of the model to the structure is just 366 pm. (B) GFP (PDB-ID 1jbz) was folded as wild-type without post-translational modification (right) and with mutations in the core that due to the sizes of their side chains would not allow folding *in vitro* (left). The "mutant" residues have elevated pLDDT (yellow and green), but overall the structure looks plausible and has low pLDDT (the walls of the β-barrel). The chromophore of GFP spontaneously forms post-translationally by condensation of three residues followed by air oxidation. The pre-condensation state cannot be trapped in a crystal structure. In this case the *in silico* model may actually represent a biologically meaningful configuration of the residues prior to forming the chromophore.



However, if a point mutation exists that has a measurable effect *in vivo* or *in vitro*, the accuracy of a model carrying that mutation may be validated: the mutation in question is introduced *in silico* into the wild-type model, scored based on energy terms, and correlated with the experimental data.

In summary, the protein structure models provided by AlphaFold2 and related algorithms surpass all previous models in terms of accuracy. Since the end of 2022, all sequences in the Uniref90 database include a corresponding structure prediction. These predicted models have been used for protein construct design, site identification for attaching probes for EPR and FRET experiments, the completion of experimental models in low-resolution X-ray electron density and cryo-EM potential density maps, and the prediction of disordered regions. The accuracy of the models is often sufficient for their use as search models in molecular replacement (Section 24.4.4.3). Taken with a grain of salt (due to the limitations of and biasing by any training set), the models are useful for biological hypothesis testing, particularly in conjunction with sparse experimental data from *e.g.* chemical footprinting, H/D exchange, or FRET. For studies requiring both high accuracy and high precision, the models are still insufficient, however. This includes confident *de novo* ligand modeling as required in structure-based drug design, the identification of potential small molecules binding to a protein cavity, and the effect of single point mutations on the structure. Future developments of machine learning algorithms in structural biology include protein dynamics, prediction of homo- and hetero-oligomeric structures of any stoichiometry and of nucleic acid complexes, plus the identification of relevant conformational states.

QUESTIONS

20.1 What is the concentration ratio between the lowest energy and next best predicted RNA secondary structure if they differ by $\Delta\Delta G$ = 5.7 kJ mol^{-1} at 37°C?

20.2 Jean-Baptiste Lamarck was a French naturalist. Assuming equal probabilities for all amino acids, what is the probability of finding the LAMARCK sequence by chance? Select the protein BLAST option at http://blast.ncbi.nlm.nih.gov/Blast.cgi and search for this sequence in proteins. How many 100% hits do you find? Select the ABC transporter. Can you find the P-loop motif for nucleotide binding? Using the ProtParam tool at http://web.expasy.org/protparam, calculate the molecular mass, the theoretical pI, and the extinction coefficients at 280 nm for reduced and oxidized Cys residues. Is the difference in extinction coefficients significant? If you do a secondary structure prediction with psipred at http://bioinf.cs.ucl.ac.uk/psipred, what secondary structure is predicted for the LAMARCK sequence? To verify the prediction, search for crystal structures of proteins using the entire sequence of the ABC transporter at www.pdb.org using the "advanced search" option and take a look at the structure of the corresponding sequence with *e.g.* Pymol or Coot. Is there a difference to the prediction?

20.3 An *ab initio* folding algorithm predicts the sequence $(K_4D_4)_n$ to adopt an α-helix. Would you need a program to come to that conclusion?

20.4 The *softmax* function converts a set of numbers of any magnitude and sign into another set with positive numbers between 0 and 1 that add up to unity. The new set can be used as probabilities in further calculations. Test this by calculating the probabilities for the set –2, –1, 1, 2, 5, 10. How probable is an event that leads to the number 5?

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{i=1}^{n} e^{z_i}}$$

20.5 PAE plots show the estimated positional error of residue *x* when model and true structure are aligned on residue *y*. Why are these plots asymmetric?

20.6 How can regions of low pLDDT help in construct design?



REFERENCES

Baek, M. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**: 871–876.
· RoseTTAFold, an alternative neural network with similar results to AlphaFold2

Coletta, A., Pinney, J. W., Solis, D. Y., Marsh, J., Pettifer, S. R. and Attwood, T. K. (2010) Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst Biol* **4**: 43.
· analysis of LCR positions within sequences

Guruprasad, K., Reddy, B. V. and Pandit, M. W. (1990) Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Eng* **4**(2): 155–161.
· generation of a table for all 400 dipeptide instability weight values (DIWV) for the calculation of the overall instability index of a protein sequence.

Hekkelman, M. L., de Vries, I., Joosten, R. P. and Perrakis, A. (2023) AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**: 205–213.
· Placing back ligands, metal ions, and prosthetic groups into AlphaFold2 models

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., … Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**: 583–589.
· Methods paper for how AF2 works

Leach, A. R. (2001) *Molecular Modelling: Principles and Applications*. Harlow and New York, Prentice Hall.
· comprehensive description of molecular modeling from a medicinal chemist's perspective

Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**(5): 911–940.
· energy parameters for RNA secondary structure predictions

Vucetic, S., Brown, C. J., Dunker, A. K. and Obradovic, Z. (2003) Flavors of protein disorder. *Proteins* **52**(4): 573–584.
· discusses different types of protein disorder

Zhang, T., Faraggi, E., Xue, B., Dunker, A. K., Uversky, V. N. and Zhou, Y. (2012) SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn* **29**(4): 799–813.
· LCR detection by a neural-network method

ONLINE RESOURCES

genomes.urv.es/OPTIMIZER
· server to optimize codon usage for mRNA

www.kazusa.or.jp/codon
· codon usage database

web.expasy.org/protparam
· calculate biophysical properties from protein sequences.

http://www.compbio.dundee.ac.uk/jpred
· predicts the secondary structure based on a multiple sequence alignment.

rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
· program rnafold for RNA secondary structure prediction.

en.wikipedia.org/wiki/List_of_RNA_structure_prediction_software
· list of RNA structure prediction software.

sparks-lab.org/SPINE-D
· SPINE D server to analyze sequences for potential disorder

www.hprd.org/PhosphoMotif_finder
· database of phosphorylation motifs



www.oxfordjournals.org/our_journals/nar/database/cat/3
· a list of DNA, RNA, and protein databases

www.uniprot.org
· UniProt database for proteins. Includes links to AlphaFold2 models, NMR, cryo-EM, and crystal structures

blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
· non-redundant nucleotide database

phylomedb.org
· database of gene phylogenies

ekhidna.biocenter.helsinki.fi/dali_lite
· DALI server

cathdb.info
· CATH database of protein folds

scop.berkeley.edu
· SCOP database of protein folds

deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology
· Deepmind's blogpost

AF2_folding_movie1 and AF2_folding_movie2
· AF2 cannot solve the Levinthal problem, *i.e.* it does not show the folding path.

alphafill.eu
· AlphaFill is a web server to complete AlphaFold2 models with missing ligands and co-factors

alphafold.ebi.ac.uk
· AlphaFold2 search engine and database

Colab
· Public implementation of the AlphaFold2 algorithmDragonFold

· Co-folding of protein sequences and ligands for protein-ligand complexes



PART IV
Methods

Parts I and II introduced the thermodynamic and kinetic concepts relevant for biochemistry. In Part III, we have connected these concepts to structure formation by and stabilization of macromolecules. Part IV introduces the biophysical techniques most frequently used to study macromolecules in solution, in crystals, and on surfaces.

21 Optical spectroscopy

Spectroscopy is based on the interaction of electromagnetic radiation with matter. Molecules may absorb energy from the incident light, which converts them into a state of higher energy, the *excited state*. From the excited state, they return to the *ground state* by releasing the energy in the form of heat or radiation. The different techniques of optical spectroscopy measure either absorption (absorption spectroscopy, Sections 21.2–21.4) or emission of radiation (emission spectroscopy, Section 21.5). Before we discuss the individual spectroscopic techniques in detail, we will begin by reviewing the properties of light and its interaction with molecules that are relevant for spectroscopy.

21.1 INTERACTION OF LIGHT AND MATTER

21.1.1 Light as an electromagnetic wave

Light is an electromagnetic wave that consists of oscillating electric and magnetic fields and propagates in a certain direction (Figure 21.1). The electric and magnetic field vectors undergo sinusoidal oscillations vertical to the propagation direction of the light. At any given time, the electric and magnetic components are vertical to each other and in phase, meaning they oscillate in synchrony, such that their maxima, minima, and zero transitions coincide.

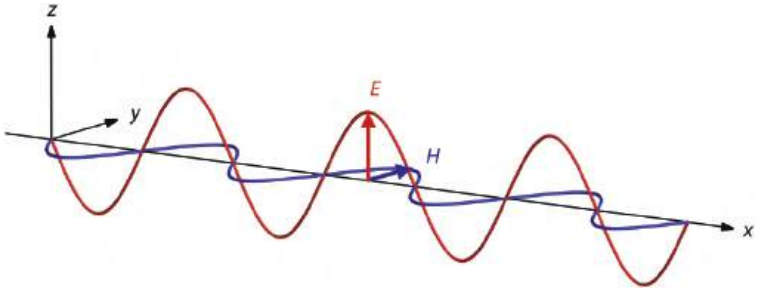


FIGURE 21.1: Electric and magnetic components of an electromagnetic wave. The electric (red) and magnetic field components (blue) of a linearly polarized electromagnetic wave propagating in the *x*-direction. The electric field vector (*E*) oscillates in the *xz*-plane, the magnetic field vector (*H*) in the *xy*-plane. *E*-field and *H*-field vectors are always perpendicular to each other.

The oscillations of the electric (E) and magnetic (H) field vectors in time (t) and space (x) are described by eq. 21.1 and eq. 21.2:

$$E(t,x) = E_0 \sin\left(2\pi\left(\nu t - \frac{x}{\lambda}\right)\right) \qquad \text{eq. 21.1}$$

E_0 is the amplitude of the electric field (in V m^{-1}), ν is the *frequency*, *i.e.* the number of oscillations of the electric and magnetic field vectors per second, and λ is the *wavelength*.

$$H(t,x) = H_0 \sin\left(2\pi\left(\nu t - \frac{x}{\lambda}\right)\right) \qquad \text{eq. 21.2}$$

H is the strength of the magnetic field (in A m^{-1}), H_0 its amplitude.

Spectroscopy is based on the interaction of molecules with the electric and/or the magnetic component of the electromagnetic wave. Most spectroscopic techniques, such as absorption, fluorescence and infrared spectroscopy, use the interaction of the molecule with the electric component. Circular dichroism (see Section 21.3.2) provides a rare example of a spectroscopic method that is based on the interaction of molecules with both the electric and magnetic components of electromagnetic radiation.

The energy E of an electromagnetic wave is proportional to its frequency ν. The proportionality constant is the Planck constant h ($6.62607 \cdot 10^{-34}$ J s):

$$E = h\nu \qquad \text{eq. 21.3}$$

The frequency of electromagnetic radiation can be expressed as the ratio of the speed of light c ($3 \cdot 10^8$ m s^{-1}) and the wavelength λ, giving :

$$E = h\frac{c}{\lambda} \qquad \text{eq. 21.4}$$

The energy of an electromagnetic wave is thus proportional to its frequency, but inversely proportional to its wavelength.

Only a small fraction of the electromagnetic spectrum (Figure 21.2) is visible to the human eye. The visible (VIS) region covers the spectral range from 400 nm (blue light) to 800 nm (red light). This range is flanked by ultraviolet (UV) and X-ray regions toward lower wavelengths (higher energy), and by infrared (IR), microwave, and radio wave regions toward higher wavelengths (lower energy).

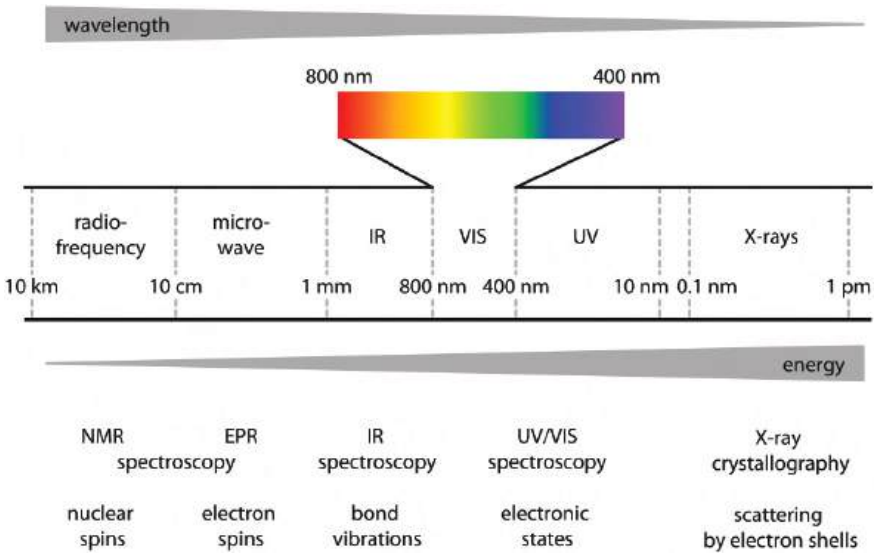


FIGURE 21.2: Electromagnetic spectrum and associated spectroscopies. The spectral region of $\lambda = 400–800$ nm is visible (VIS) for the human eye. Infrared radiation (IR), microwaves, and radio waves are of lower energy (higher wavelength), ultraviolet (UV) radiation and X-rays are of higher energy (lower wavelength). Nuclear magnetic resonance (NMR) spectroscopy induces transitions of nuclear spins (Section 22.1), electron paramagnetic resonance (EPR) spectroscopy induces transitions of electron spins (Section 22.2). IR spectroscopy excites bond vibrations (Section 21.4), UV/VIS spectroscopy (Sections 21.2 and 21.3) induces transitions between electronic states of molecules. X-rays are scattered by the electron shell of molecules, a process that does not involve absorption, but is important for structure determination (Chapter 24).

21.1.2 Principles of spectroscopy: transitions in two-state systems

Spectroscopy is based on transitions between different energy levels of a molecule that are induced by the energy of the incident electromagnetic radiation. Such states can be different electronic states (UV/VIS absorption and fluorescence spectroscopy; Sections 21.2, 21.3, and 21.5), vibrational states (IR spectroscopy; Section 21.4), rotational states, or spin states of nuclei (NMR; Section 22.1) or electrons (EPR; Section 22.2). The energy of these states is quantized: only defined states can be populated, leading to a set of discrete energy levels.

Incident light can provide the energy to induce transitions from the lower energy level, the ground state, to a higher energy level, the excited state. To do so, the energy E of the electromagnetic wave (eq. 21.3, eq. 21.4) has to match the energy difference ΔE between the two states involved (Figure 21.3). It is important to note that the energy difference ΔE refers to the difference between the two states in a single molecule, resulting in very small numbers. Electronic transitions can typically be induced by UV/VIS light in the range of 200–800 nm, corresponding to energies on the order of 10^{-19} J. Sometimes these energies are given in electron volts (eV). An electron volt corresponds to the work associated with moving an electron (with the elementary charge of $1.602 \cdot 10^{-19}$ C) against a potential difference of 1 V. One electron volt thus equals $1.602 \cdot 10^{-19}$ J. Alternatively, molar energies are given: 10^{-19} J per molecule add up to 16 kJ mol^{-1}.

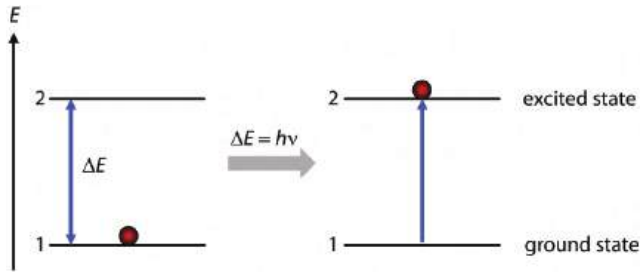


FIGURE 21.3: Light-induced transitions in a two-state system. A molecule in its lower energy ground state (1, left) can undergo a transition to the excited state (2, right, blue arrow) upon irradiation with light of the frequency ν if the energy of the photons ($h\nu$) equals the energy difference ΔE that separates the two states.

The population of the two (electronic or other) states in the absence of irradiation depends on the energy difference ΔE separating the two states and the *thermal energy* of the molecules. For single molecules, the thermal energy is the product of the Boltzmann constant k_B and the temperature T:

$$E = k_B T \qquad \text{eq. 21.5}$$

For one mol of molecules, the thermal energy is

$$E_{molar} = E \cdot N_A = N_A k_B T = RT \qquad \text{eq. 21.6}$$

N_A is the Avogadro constant, R the general gas constant (Section 1.5.5, Box 1.8). At room temperature (298 K), the thermal energy of a molecule, $k_B T$, is ~$4.1 \cdot 10^{-21}$ J, the thermal energy of one mol of molecules, RT, is ~2.5 kJ mol^{-1}.

The population of two states separated by ΔE is described by the *Boltzmann distribution* (eq. 21.7; see Section 3.1). The probability p_2 that a molecule is in the high-energy state 2 relative to the probability p_1 that it is in the lower energy state 1 is

$$\frac{p_2}{p_1} = \exp\left(-\frac{\Delta E}{k_B T}\right) \qquad \text{eq. 21.7}$$

For an ensemble of molecules, the number of molecules n_2 in the higher energy state relative to the molecules n_1 in the lower energy state is

$$\frac{n_2}{n_1} = \exp\left(-\frac{\Delta E_{molar}}{RT}\right) \qquad \text{eq. 21.8}$$

The higher the energy difference between states 1 and 2, the lower the probability that the molecules are in the high-energy state. The higher the temperature and the thermal energy of the molecules, the higher the probability that the molecules populate the high-energy state (see Sections 3.1, 3.2).

The number of light-induced transitions from state 1 to 2 (1→2) is determined by the product of the population of state 1 and the transition probability (1→2). The incident light also induces transitions from state 2 to 1. The number of these 2→1 transitions is related to the product of the population of state 2 and the transition probability (2→1). For a simple two-state system, the probabilities for a transition 1→2 and for the reverse transition 2→1 are equal. At the beginning of the illumination, state 1 is more populated than state 2, and by irradiation with light we will be able to induce more transitions from 1→2 than from 2→1. With increasing numbers of molecules in state 2, reverse transitions will start to pick up. Eventually, we will reach equal populations of states 1 and 2. Now the numbers of transitions 1→2 and 2→1 are equal: the transition is saturated, and absorption is no longer observed (Figure 21.4).

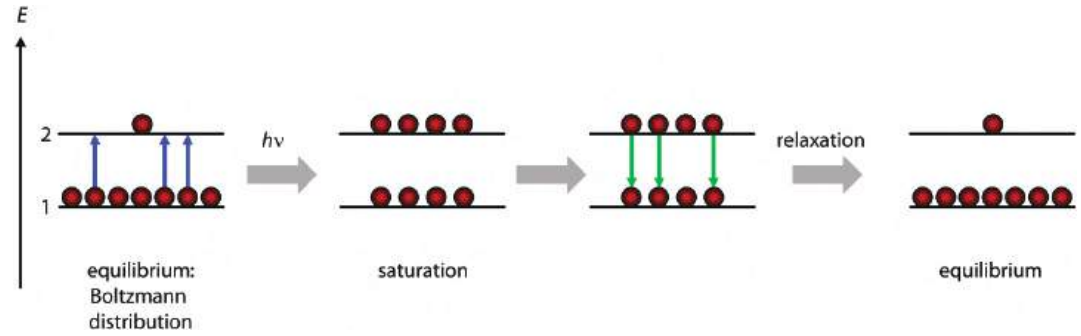


FIGURE 21.4: Transitions in a two-state system: Boltzmann distribution, excitation, saturation, and relaxation. In equilibrium, the lower energy ground state (1) will be more populated than the excited state (2) according to the Boltzmann distribution (eq. 21.7, eq. 21.8). Irradiation with light leads to transitions from the ground state to the excited state (blue arrows). The transition is saturated when both states are equally populated. In the absence of light, molecules will relax from the excited state to the ground state (green arrows), restoring the equilibrium distribution.

21.2 ABSORPTION

21.2.1 Electronic, vibronic, and rotational energy levels

As outlined in Section 21.1, a molecule can absorb light when the energy of a photon corresponds to the energy separating its *electronic ground state* (S_0) from its *electronically excited state* S_1 (or S_2, or higher S_n states). The "S" refers to singlet states, in which electron spins (see Section 22.2) are paired. Absorption of light converts the absorbing molecule into the excited state, in which electrons are more separated from the nuclei compared to the ground state. The energy of the absorbed light is required to overcome the electrostatic attraction that counteracts the separation of the electrons from the nuclei. In addition to electronic excitation, electromagnetic radiation can also induce transitions between *vibronic or rotational states* of the molecule. The energy differences between vibronic states are about one order of magnitude smaller than for electronic states, and the differences between adjacent rotational states are about two orders of magnitude smaller. The electronic, vibronic, and rotational energy levels of a molecule can be depicted schematically in an energy diagram, the *Jablonski diagram* (Figure 21.5). Such a diagram is useful to illustrate processes occurring in spectroscopy, and we will come back to this simplified, but clear depiction several times in the following.

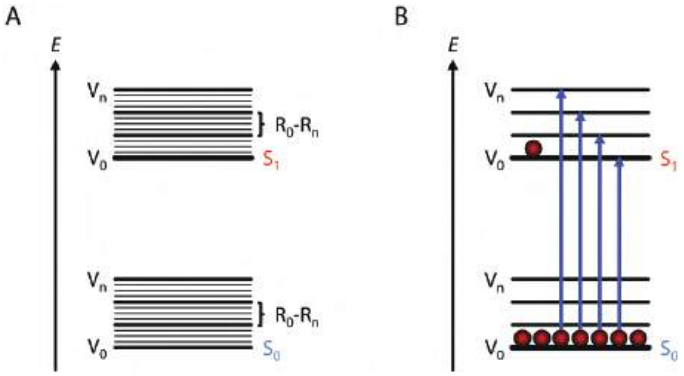


FIGURE 21.5: Jablonski diagram. A: A molecule can exist in two electronic states, S_0 and S_1 (thick lines). S_0 is the ground state, corresponding to the lowest energy electron configuration. S_1 is the excited state in which electrons are more separated from the nuclei compared to the ground state. The S_0 and S_1 states are subdivided into different vibrational states, V_0-V_n (medium lines) and rotational states (R_0-R_n, thin lines). The energy differences between adjacent vibronic energy levels are about one order of magnitude smaller than the energy differences between electronic states; the energy differences between neighboring rotational energy levels are two orders of magnitudes smaller. B: According to the Boltzmann distribution, spectroscopy at room temperature starts from S_0,V_0. Irradiation with light can induce transitions (blue arrows) from S_0,V_0 to different vibronic levels of S_1.

At room temperature (298 K), the thermal energy RT is about 2.5 kJ mol^{-1} (eq. 21.6). This value is on the order of energetic differences between rotational energy levels. Transitions between rotational states can therefore be excited by thermal energy. In contrast, the thermal energy is much smaller than the energy differences between vibronic or electronic states: transitions between vibronic or electronic states cannot be induced by thermal energy at ambient temperatures. Spectroscopy at room temperature therefore starts from the lowest electronic and vibronic energy levels of the molecule, S_0,V_0 (Figure 21.5; see Section 3.2), but not from the lowest rotational level.

21.2.2 Transitions and transition dipoles

Generally, transitions between two states of a molecule can be induced by perturbations, *i.e.* by external electric or magnetic fields. The effectiveness of the perturbation to induce a transition depends on the extent to which it can deform the initial state such that it resembles the final state. Incident light induces transitions between electronic states by altering the charge distribution of the molecule. This change in charge distribution is connected to an induced dipole in the molecule. According to classical mechanics, the dipole $\boldsymbol{\mu}$ induced by an electric field E depends on the polarizability α of the molecule:

$$\boldsymbol{\mu} = \alpha \cdot \boldsymbol{E} \qquad \text{eq. 21.9}$$

The *transition dipole* $\boldsymbol{\mu}$ quantifies the direction and magnitude of changes in the charge distribution within a molecule. The direction of the dipole moment is a property of the molecule, and is fixed with respect to its geometry. The transition dipole moment is a measure of the capacity of the incident light to distort the initial state toward the final state. Therefore, the magnitude of the transition dipole $|\boldsymbol{\mu}| = \mu$ determines the probability for a transition between these states upon irradiation. Transitions with $\mu \neq 0$ are called *allowed transitions*, whereas transitions associated with a transition dipole of zero are *forbidden*. Although they can be observed experimentally, the transition probabilities and absorption intensities are very low. Selection rules tell us if a particular transition between two states is allowed in a quantum-mechanical system.

In Section 21.1.2, we have considered the behavior of two-state systems independent of the molecular properties. Molecules consist of atoms that are linked by covalent bonds. The dependence of the potential energy of two atoms on their distance is described by the *Lennard-Jones potential* (Figure 17.13; see Section 17.2.2) that reflects the balance of the attractive and repulsive forces between the two atoms. At the distance r_{min}, the van der Waals distance of the two atoms, the potential energy reaches a minimum. In electronically excited states, electrons usually occupy non-bonding and anti-bonding orbitals, resulting in less electron density between the nuclei and higher repulsion forces between their positive charges, and an increased bond length r_{min} in the electronically excited state compared to the ground state. The electronic ground state and the first electronically excited state of a molecule can therefore be depicted as two Lennard-Jones potential curves, shifted along the r-axis (Figure 21.6). Within these potential energy curves, we can now include the vibronic levels as horizontal lines for ground state and excited state. Absorption is an inherently rapid process, and occurs on the time scale of 10^{-15} s. This time scale is too short for bonds to vibrate or for conformational changes of the absorbing molecule to occur. Hence, the molecule is static on the time scale of absorption, and bond lengths do not change. As a consequence, absorption can be depicted as a vertical transition in Lennard-Jones potential diagrams (Figure 21.6). This principle is known as the *Franck-Condon principle*, or the *principle of vertical transitions*.

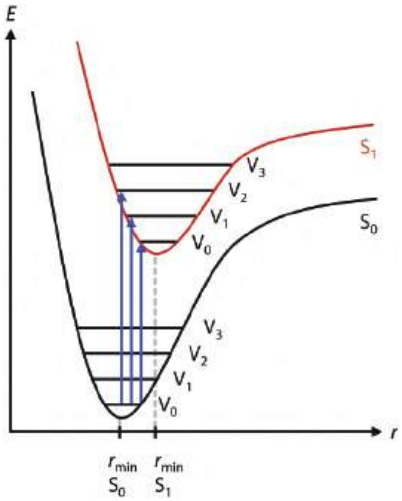


FIGURE 21.6: Vertical transitions in the Lennard-Jones potential. On the time scale of absorption, the geometry of molecules does not change, and bond lengths remain constant. As a consequence, transitions between the electronic ground state S_0 and the excited state S_1 can be depicted as vertical arrows (blue) in the Lennard-Jones potential. The vibronic levels V_0-V_3 for both states are indicated.

At room temperature, absorption of light causes transitions from the electronic and vibronic ground state, S_0,V_0 (see Section 21.1.2), to the electronically excited state S_1. Depending on the energy of the excitation light, the molecules can be excited to higher vibronic levels V_x of the S_1 state. An *absorption spectrum* is the dependence of absorption by a molecule, a *chromophore*, as a function of the wavelength of the irradiating light (Figure 21.7). Such a spectrum typically consists of a set of lines. Each set corresponds to a particular $S_n \rightarrow S_{n+1}$ transition, and the individual lines of the set correspond to the individual vibronic transitions involved. In solution, the energies involved in these transitions of the dissolved molecule, the solute, show some variation due to interactions with the solvent. Hence, the measured absorption spectrum of chromophores in solution is the envelope of all vibronic transitions of one set of electronic transitions. The lines underneath the broad absorption bands are called the *fine structure* of the spectrum.

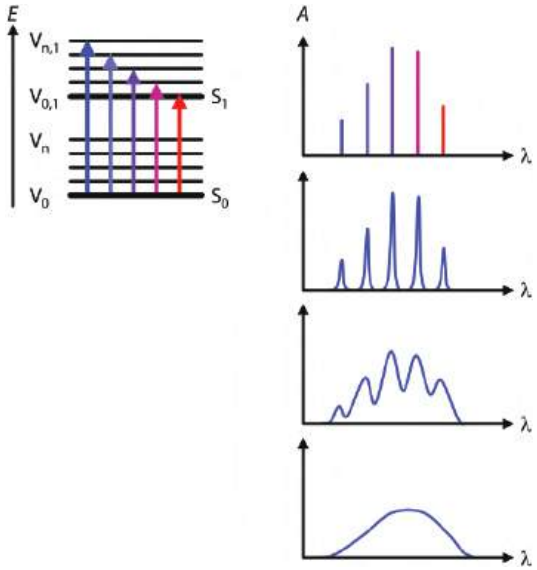


FIGURE 21.7: From line spectra to continuous spectra. Left: Jablonski diagram for a molecule that can undergo transitions from S_0,V_0 to different vibrational levels of S_1 (arrows). The transitions in the Jablonski diagram predict a set of absorption lines at the wavelength corresponding to the energy of each transition (right, top). Such a line spectrum is observed in the absence of collisions with other molecules or with solvent molecules. In the presence of collisions, the discrete energy levels become more degenerate, and the lines in the spectrum broaden and overlap until a continuous spectrum is observed (right, bottom).

An absorption band is characterized by the wavelength of maximum absorption (λ_{max}) and the *extinction coefficient* ε_λ (in $M^{-1}\,cm^{-1}$) at this wavelength (see Section 21.2.3), which is a measure of the absorption intensity. The absolute absorption intensity, *i.e.* the area under the absorption band, is determined by the *dipole strength D*. The dipole strength $D_{i\to f}$ (in the unit debye[2], with 1 debye = $3.336\cdot10^{-30}$ C·m) of the transition from the initial state *i* to the final state *f* is related to the measured spectrum by eq. 21.10:

$$D_{i\to f} = 9.18\cdot10^{-3}\int \frac{\varepsilon}{\nu}\,d\nu \qquad \text{eq. 21.10}$$

Thus, the dipole strength can be obtained by integration of the absorption band in a spectrum that has been converted into the molar extinction coefficient ε (see Section 21.2.3) as a function of the frequency ν. The dipole strength is proportional to the square of the transition dipole.

21.2.3 The Lambert-Beer law

One of the most important applications of absorption spectroscopy is the determination of concentrations. To derive a relationship between absorption and concentration, we consider a solution containing absorbing molecules (Figure 21.8) that is illuminated by light of the intensity I_0. The light path through the sample can be divided into slices with the incremental thickness d*x*. The light entering one of these slices will have the intensity *I*. The intensity of the light exiting from the slice, the transmitted light, will be reduced to *I*–d*I* due to absorption by the molecules within this slice. The change in intensity d*I* is proportional to the intensity *I* of the incident light, and to the concentration *c* of the absorbing molecules. We name the proportionality constant *a*. This constant is a measure of the absorption strength of the molecule. The reduction in light intensity due to absorption is accounted for by the negative sign in eq. 21.11:

$$dI = -a\cdot c\cdot I\,dx \qquad \text{eq. 21.11}$$

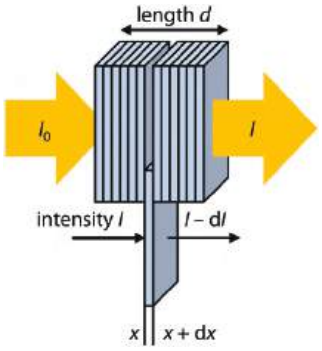


FIGURE 21.8: Lambert-Beer law. A sample in a cuvette of path length *d* is illuminated by light. Within a slice of the depth d*x* (bottom) absorption leads to a reduction of the light intensity from *I* (intensity of the incident light) to *I*–d*I* (intensity of the transmitted light). By integrating over the path length *d* and the light intensity, the Lambert-Beer law can be derived.

Separation of variables in eq. 21.11 gives

$$\frac{dI}{I} = -a\cdot c\,dx \qquad \text{eq. 21.12}$$

Eq. 21.12 can then be integrated over all slices of the sample, from $x = 0$ to the path length *d* and from the initial light intensity I_0 to the intensity *I* exiting the sample:

$$\int_{I_0}^{I} \frac{\mathrm{d}I}{I} = -a \cdot \int_{0}^{d} c \, \mathrm{d}x \qquad \text{eq. 21.13}$$

The concentration c of the absorbing molecule is homogeneous over the sample. Therefore, c does not depend on the location of x, and can be considered as constant. We can evaluate the integrals to

$$\ln \frac{I}{I_0} = -a \cdot c \cdot d \qquad \text{eq. 21.14}$$

Conversion of the natural logarithm in eq. 21.14 into the decadic logarithm (see Section 31.5) according to

$$\ln \frac{I}{I_0} = \ln 10 \cdot \log \frac{I}{I_0} = -\ln 10 \cdot \log \frac{I_0}{I} \qquad \text{eq. 21.15}$$

and definition of the constant $a/\ln 10$ as the molar *extinction coefficient* ε gives

$$\log \frac{I_0}{I} = \varepsilon c d \qquad \text{eq. 21.16}$$

The logarithm of the intensity of the incident light I_0, divided by the intensity of the transmitted light I, is called the absorption A. We thereby obtain the *Lambert-Beer law* as

$$A = \varepsilon c d \qquad \text{eq. 21.17}$$

The Lambert-Beer law is central for the photometric determination of the concentration of absorbing molecules. If the extinction coefficient for a molecule of interest is known, we can calculate its concentration in solution from the measured absorbance (see Section 21.2.7). It is important to keep in mind that absorption is defined on a logarithmic scale (see Section 21.3.6). Also, the Lambert-Beer law holds only for ideal solutions, in which the chromophores do not interact. At high chromophore concentrations, the linear relationship between absorption and chromophore concentration breaks down due to physical and chemical interactions between molecules (Section 21.2.8).

21.2.4 Solvent effects and influence of the local environment

Absorption is linked to a change in charge distribution: the electronically excited state has a different charge distribution than the ground state. In solution, the configuration of solvent molecules around the solute will be such that the solvent stabilizes the electronic state of the solute molecule. This effect is particularly important for aqueous solutions because water molecules are strong permanent dipoles. Ordering of water molecules around aromatic moieties is the molecular basis for the hydrophobic effect (Section 18.5.1 and Figure 18.41). Charged solutes are stabilized in aqueous solutions by hydration, an energetically favorable orientation of the surrounding water molecules (Figure 21.9). When the solute molecule absorbs light and undergoes a vertical transition (Figure 21.6) to the electronically excited state, its charge distribution changes within 10^{-15} s. On this time scale, not only the chromophore and its geometry, but also the surrounding water molecules are unchanged. However, the arrangement of water molecules that was optimal for the stabilization of the ground state is not the energetically favorable geometry for hydration of the chromophore in its excited state. Therefore, this geometry of water molecules now destabilizes the excited state relative to a non-polar solvent, and increases its energy (Figure 21.9). Consequently, the energy difference between S_0 and S_1 levels is increased compared to a non-polar solvent, and the absorption wavelength will be changed accordingly. A shift in the absorption wavelength toward the blue end of the spectrum (higher energy) is called a *hypsochromic effect*, a shift toward the red (lower energy) is called a *bathochromic effect*.

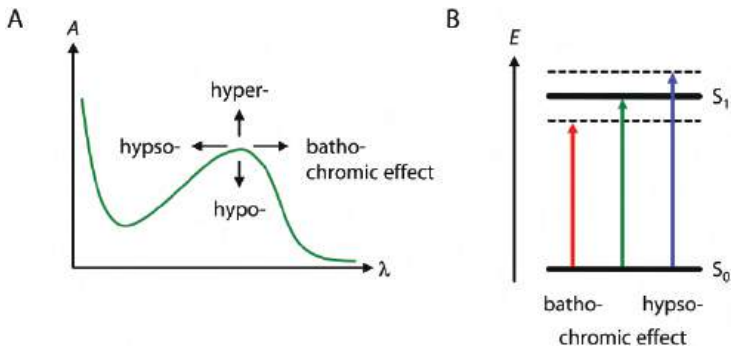


FIGURE 21.9: Solvent effects on electronic transitions. A: The environment of a chromophore affects its absorption spectrum, and can shift the maximum of the absorption to lower wavelengths (blueshift, hypsochromic effect) or to longer wavelengths (redshift, bathochromic effect). An increase in the extinction coefficient corresponds to a hyperchromic effect, a decrease to a hypochromic effect. B: Energy diagram for batho- and hypsochromic effects of solvents.

In contrast to changes in the absorption wavelength, changes in the extinction coefficient are called *hypochromic* (decrease) or *hyperchromic effects* (increase; Figure 21.9). These effects are caused by effects of the solvent on the transition dipole and thus on the dipole strength (eq. 21.10) of the molecule.

Similar to the effects of the solvent on electronic transitions of a solute molecule, the local environment of a chromophoric group within a larger molecule such as a protein or a nucleic acid influences its absorption. For example, the absorption of chromophoric side chains of amino acids depends on their chemical environment within the three-dimensional structure of the protein. Therefore, the absorption of a protein in the folded state under native conditions differs from its absorption under denaturing conditions, where it is unfolded and all side chains are solvent-exposed. Absorbance is therefore a powerful spectroscopic probe to monitor protein folding and unfolding (see Section 21.2.7). The direction of the change cannot be predicted, and hyper- or hypochromic as well as batho- and hypsochromic effects upon folding are observed.

The absorption of the nucleobases in the DNA also depends on the environment. The absorbance of the bases in the DNA is influenced by their neighbors within the same strand, and by their pairing partners on the opposite (complementary) strand. DNA shows a strong hyperchromic effect upon dissociation of the double helix into single strands, and absorption is therefore routinely used as a sensitive spectroscopic probe to monitor DNA melting (see Section 21.2.7).

21.2.5 Instrumentation

Absorption is measured in spectrometers that consist of a light source, typically a deuterium arc lamp for UV- and tungsten lamp for visible light or a Xenon arc lamp (UV/VIS), a monochromator to select a certain wavelength of the incident light, a cuvette holder, and a photomultiplier that detects the light transmitted by the sample (Figure 21.10). The photons that enter the photomultiplier generate electrons by the photoelectric effect. These secondary electrons are accelerated in an electric field and amplified to a measurable photocurrent that is proportional to the light intensity. This simplest type of absorption spectrometer is termed single-beam photometer. The absorption of the sample, typically a buffered solution of a protein or nucleic acid, is calculated from the intensities of the incident and the transmitted light. Solvent contributions to the absorption of the sample need to be measured in a second measurement of a sample with the solvent only, and are subtracted from the absorption of the sample. If the wavelength of the incident light is varied, an *absorption spectrum*, $A(\lambda)$, is obtained.

In a double-beam photometer, the incident light is split before the sample. One part is passed through the sample, and the second part is passed through a reference cuvette containing the solvent only. The difference in absorption between sample and reference cell then directly reflects the absorption of the solute in the sample cell. Using a double-beam photometer, *difference spectra* may be measured directly. In this case, both cuvettes may contain samples, *e.g.* a protein in the native state in one cuvette and the same protein in the denatured state in the second cuvette.

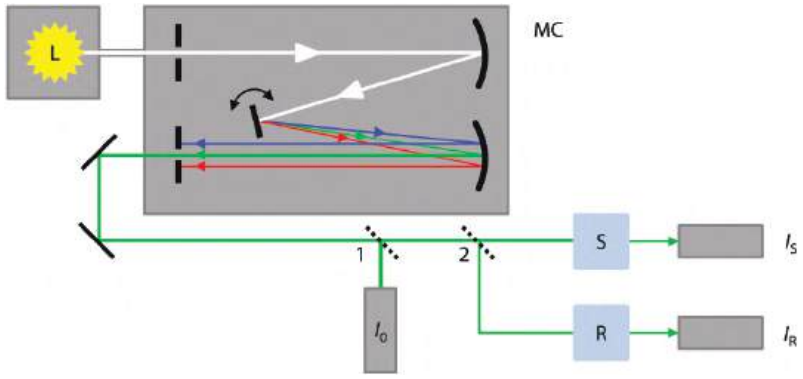


FIGURE 21.10: Single- and double-beam photometer. A light source (L, yellow) emits white light. The monochromator (MC) selects the desired wavelength. The selection depends on the position of an adjustable diffraction grating (double-headed arrow). A small fraction of the light exiting the monochromator is reflected by a partially transmitting mirror (1) to measure the intensity of the incident light I_0. The remaining light passes the sample (S) in a cuvette. The transmitted light I_S is detected by a photomultiplier. In a double-beam spectrometer, the light is split a second time (2). One part passes the sample, the second a reference cuvette (R) that contains solvent only. The absorption of the solute molecule in the sample is then calculated from the difference in the intensities of transmitted light, I_S–I_R.

21.2.6 Biological chromophores

Proteins typically show absorption in the UV range, caused predominantly by the aromatic side chains of Tyr and Trp due to $\pi \rightarrow \pi^*$ transitions (Figure 21.11). Tyr and Trp are not very common amino acids, with a relative abundance of 3.2% and 1.3%, respectively, and proteins therefore only contain a small number of chromophores. The extinction coefficients at 280 nm for Tyr and Trp are around 1200 and 5600 M^{-1} cm^{-1}. Both Tyr and Trp also exhibit weak absorption below 220 nm. Despite the lower extinction coefficient, tyrosine residues contribute significantly to protein absorption due to their higher abundance. An average protein of 100 residues would thus typically have an extinction coefficient on the order of 10000 M^{-1} cm^{-1}. Although Phe contributes to protein absorption near 260 nm because of a symmetry-forbidden $\pi \rightarrow \pi^*$ transition ($\varepsilon \approx 400$ M^{-1} cm^{-1}), this weak band is usually not observed experimentally in proteins that also contain Trp and Tyr.

The peptide bond shows two absorption bands in the UV range, one around 210–220 nm as the result of a forbidden $n \rightarrow \pi^*$ transition ($\varepsilon \approx 100$ M^{-1} cm^{-1}), and a much stronger band at around 190 nm ($\varepsilon \approx 7000$ M^{-1} cm^{-1}) because of a $\pi \rightarrow \pi^*$ transition. The side chains of some amino acids, such as Asp, Glu, Asn, Gln, Arg, His, and Cys, and Tyr and Trp, absorb in the same spectral range. Their contribution can usually not be measured in the context of proteins, however, because it is very small, and because these residues are always outnumbered by the peptide bonds. Disulfides (cystines) show red-shifted absorption at 250–270 nm ($\varepsilon \approx 300$ M^{-1} cm^{-1}), which is usually too small to be exploited in absorption measurements.

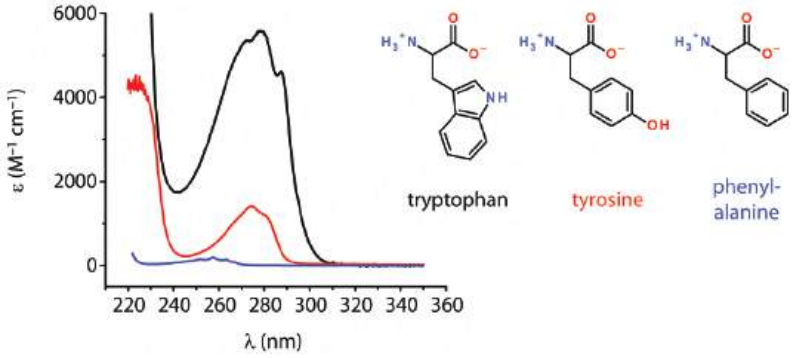


FIGURE 21.11: Absorption of aromatic amino acids. Absorption spectra of tryptophan (black) tyrosine (red) and phenylalanine (blue), converted to the molar extinction coefficient. Dixon *et al.* (2005) Photochem. Photobiol. 81(1): 212–213. The structures of tryptophan, tyrosine, and phenylalanine are depicted on the right.

The nucleobases of DNA and RNA absorb strongly in the UV as a result of several n→π* and π→π* transitions with maxima between 255 and 275 nm (Figure 21.12). The extinction coefficients at 260 nm are about 8000 M^{-1} cm^{-1} for pyrimidines, and 15000 M^{-1} cm^{-1} for purines. The absorbance spectrum of DNA shows a broad maximum at around 260 nm, with an average extinction coefficient of about 10000 M^{-1} cm^{-1} per nucleotide. In contrast to proteins, each building block provides a chromophore, and the average extinction coefficients for nucleic acid molecules are very high. Therefore, absorption at 260 nm provides a highly sensitive method to determine nucleic acid concentrations.

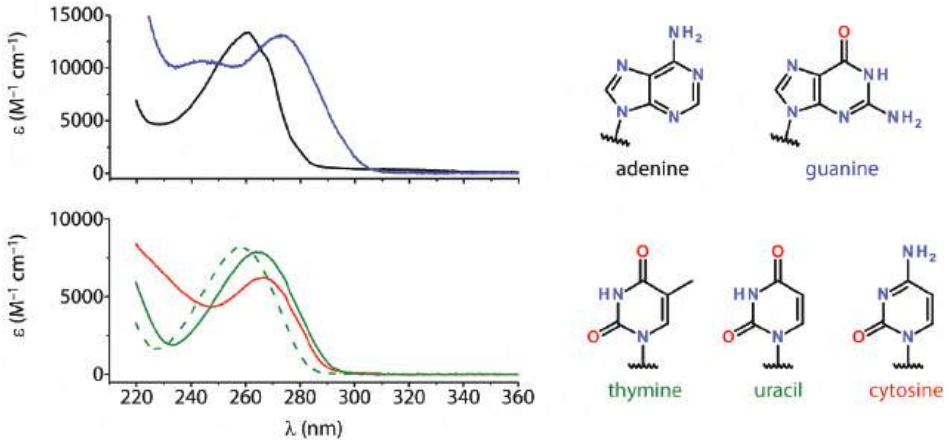


FIGURE 21.12: Absorption of bases in nucleic acids. Absorption spectra of adenine (black) and guanine (blue, top), and cytosine (red), thymine (green), and uracil (green, dotted line, bottom). All spectra are normalized to the molar extinction coefficient. Dixon *et al.* (2005) Photochem. Photobiol. 81(1): 212–213. The structures of the nucleobases are depicted on the right.

Several cofactors and prosthetic groups also absorb light in the UV and visible regions (Figure 21.13). Examples include NADH, an electron donor in redox reactions (Chapter 5 and Section 5.6) that absorbs at 340 nm due to an n→π* transition of the conjugated CO group in the nicotinamide ring, and flavins (FMN, FAD) that absorb at 445–465 nm. Flavonoids, such as anthocyanins and flavones, are responsible for the yellow, red, and blue colors of flowers and fruits. Tetrapyrroles are widespread chromophores of four pyrrole rings that can either form a macrocyclic, conjugated ring system such as in the porphyrins, or linear structures, such as in the bile pigment biliverdin. Heme, the prosthetic groups in hemoglobin, myoglobin, and cytochromes, is a redox-active porphyrin-Fe^{2+}/Fe^{3+} complex. Their intense absorption at 400–500 nm, the "Soret" band, is caused by a π→π* transition in the extensive aromatic system with electron delocalization over all four pyrrole rings.

The photosynthesis pigment chlorophyll (Section 5.6.3) is a Mg^{2+} complex of a modified porphyrin. Carotenoids (carotenes and xanthophylls; Figure 21.13) are also involved in photosynthesis and show strong absorption between 450 and 550 nm. The absorption of UV light (at 335 nm) by melanin, generated by the oxidation and polymerization of tyrosine residues, protects our skin from UV damage and cancer.

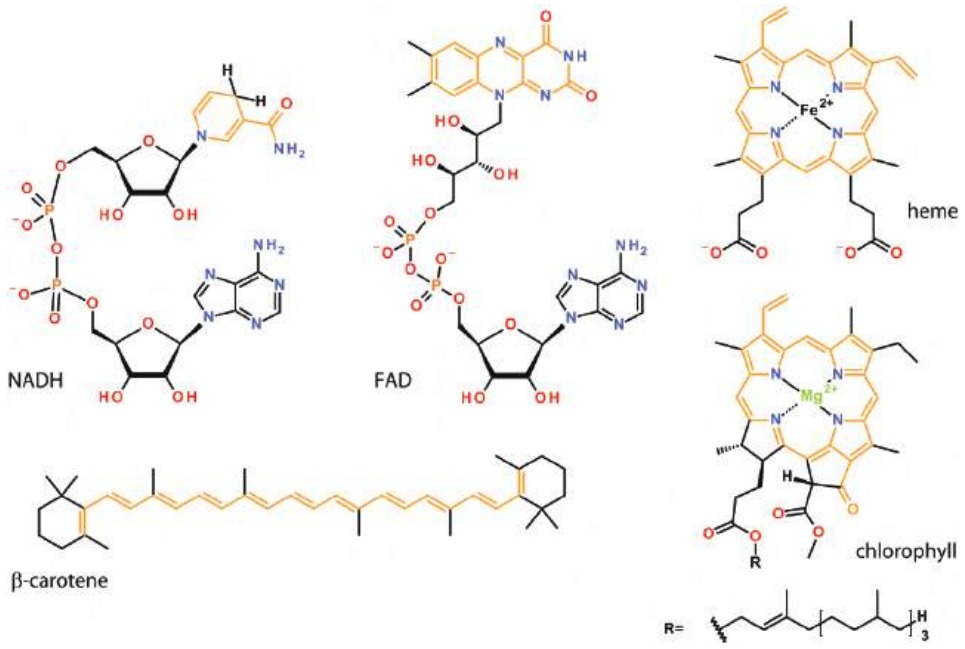


FIGURE 21.13: Cofactors and prosthetic groups that show UV absorption. Nicotinamide adenine dinucleotide (NADH; reduced form), flavin adenine dinucleotide (FAD; oxidized form), the porphyrins heme and chlorophyll, and the carotenoid β-carotene. The delocalized π-electron system responsible for the absorption is highlighted in orange.

Vitamin A, the precursor of the vision pigment retinal, and retinal itself absorb at 370 nm (Figure 21.14). The absorption of retinal is shifted to 500 nm in complex with the protein opsin. It further changes during the *cis*→*trans* isomerization (Section 17.1) of retinal, the central reaction in the vision process (Box 21.1). Retinal is also the chromophoric unit in light-driven proton pumps and ion channels (bacteriorhodopsin, channelrhodopsin, and halorhodopsin; Box 21.2, see Section 4.3).

BOX 21.1: THE VISION PROCESS.

Retinal (Figure 21.14; λ_{max} = 370 nm) is the central chromophore in our vision process. Upon binding of 11-*cis* retinal to the seven-helix transmembrane protein opsin, the absorption maximum undergoes a huge bathochromic shift to λ_{max} = 500 nm. 11-*cis* retinal has an extinction coefficient of ε = 40000 M^{-1} cm^{-1}. When 11-*cis* retinal absorbs light, it isomerizes to all-*trans* retinal. The affinity of opsin for all-*trans* retinal is low, and retinal dissociates from rhodopsin. Opsin is a G-protein coupled receptor that interacts with the G-protein transducin. Transducin is a heterotrimeric G-protein. Upon activation by opsin, its α-subunit exchanges bound GDP for GTP and activates a phosphodiesterase that hydrolyzes cyclic GMP (cGMP) to GMP. cGMP maintains ligand-gated sodium channels in the plasma membrane in the open form. Its hydrolysis accordingly leads to closing of these channels. The resulting hyperpolarization (see Section 5.7) of the rod cells affects the transmission of signals from the rod cells to bipolar cells to the ganglia of the optical nerve.

(*Continued*)

Absorption of a single photon leads to isomerization of one retinal molecule, but each opsin activates many transducins and leads to opening of numerous sodium channels. This signal amplification renders our eyes very sensitive optical instruments!

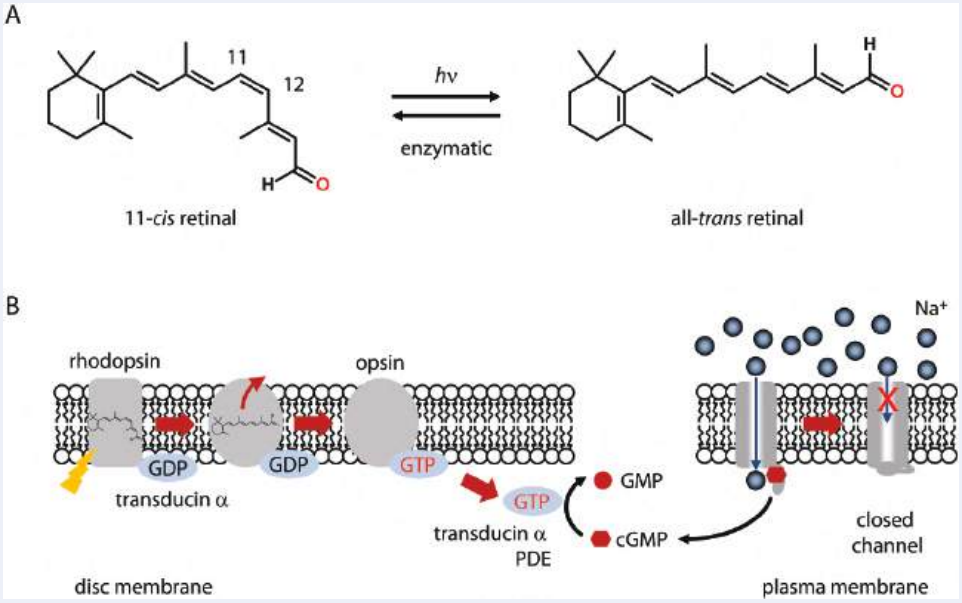


FIGURE 21.14: Retinal, rhodopsin, and the vision process. A: 11-*cis* retinal is converted to all-*trans* retinal upon illumination (*hν*). All-*trans* retinal is enzymatically regenerated to 11-*cis* retinal. B: 11-*cis* retinal is bound to the protein opsin, an integral membrane protein in the disc membrane of rod cells. Illumination causes isomerization of the retinal to the all-*trans* form that dissociates from opsin (red arrow). The resulting activated opsin converts the α-subunit of the G-protein transducin into the GTP form that has phosphodiesterase (PDE) activity. The PDE hydrolyzes cyclic GMP (cGMP) to GMP, leading to the closure of cGMP-gated sodium channels in the plasma membrane and hyperpolarization of the rod cell.

BOX 21.2: LIGHT-DRIVEN PROTON PUMPS AND LIGHT-GATED ION CHANNELS.

Bacteriorhodopsin is a transmembrane protein in the purple membrane of archaea that acts as a light-driven proton pump (Figure 21.15). Its chromophore all-*trans* retinal is covalently bound to a lysine *via* a protonated Schiff base, and separates the pore into two half-channels. Upon absorption of light (λ_{max} = 568 nm), retinal isomerizes to 13-*cis* retinal. The resulting conformational change reduces the polarity around the Schiff base, leading to deprotonation and release of the proton to the outside of the cell. A subsequent conformational change in the photocycle of bacteriorhodopsin restores the polarity around the Schiff base, which becomes reprotonated from the inside. One absorption event is thus coupled to the net transport of one proton from the inside of the bacterium to the outside. The light energy is thereby converted into a proton-motive force across the plasma membrane (Section 4.2) that can be exploited to drive the synthesis of ATP. Halorhodopsin is a light-gated chloride pump in halobacteria. Both bacteriorhodopsin and halorhodopsin use light as the energy source for active transport. Channelrhodopsins are light-gated ion channels (Section 4.3) in algae. The discovery of channelrhodopsins and their application for the manipulation of membrane potentials in living cells and animals has opened up the field of optogenetics, and has provided a huge impulse to the field of neurophysiology.

(*Continued*)

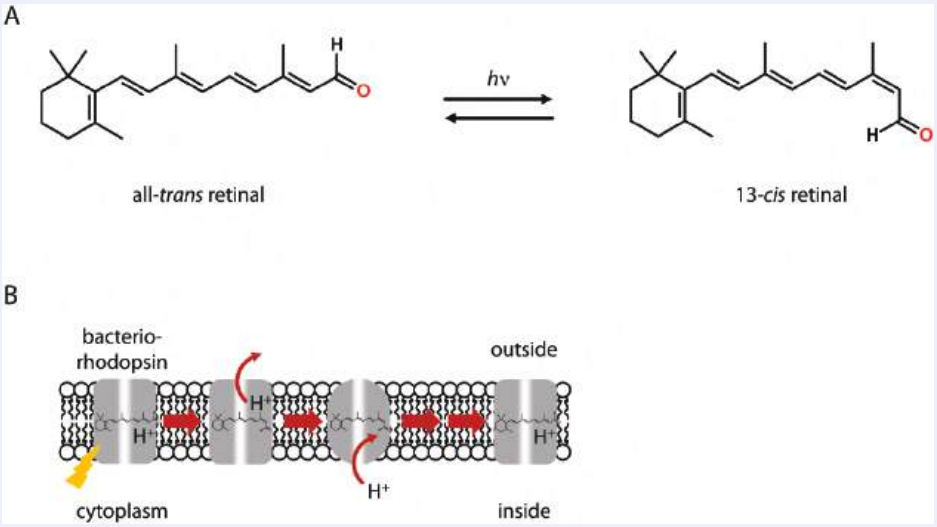


FIGURE 21.15: Light-gated ion pumps and channels. A: Light-driven ion pumps use the chromophore all-*trans* retinal that isomerizes to 13-*cis* retinal upon absorption of light. B: Bacteriorhodopsin is a light-driven proton pump that uses the energy of light to generate a proton-motive force (see Section 4.2). The chromophore all-*trans* retinal is bound to the protein by a protonated Schiff base and blocks the proton channel. Upon absorption and isomerization of retinal, and a conformational change of the protein, the Schiff base becomes deprotonated. A proton is released to the outside. Regeneration of the original state of bacteriorhodopsin, *i.e.* the all-*trans* form of retinal and the original conformation of the opsin, are coupled to the uptake of a proton from the cytoplasm. Overall, one absorption event is coupled to transport of one proton from the cytoplasm to the outside. Halorhodopsin and channelrhodopsins use a similar mechanism to couple the transport of chloride and other ions to the absorption of light.

21.2.7 Applications

21.2.7.1 Concentration determination

According to the Lambert-Beer law (eq. 21.17), absorption is proportional to chromophore concentration. Absorption is therefore a useful tool to determine concentrations in solution, provided the extinction coefficient of the molecule of interest in the solvent used is known. In the case of proteins and DNA, the effect of the local environment has to be taken into account. For an unfolded protein, with solvent-exposed side chains, the extinction coefficient at 280 nm can be calculated by simply adding the extinction coefficients for all chromophoric side chains present. Typically, the aromatic side chains of Tyr and Trp have to be taken into account; the much smaller contribution of cystine groups (cysteines forming disulfide bonds) may also be considered. The absorption differences between denatured and native states of a protein typically do not exceed 10% (Figure 21.19). As a consequence, the calculated ε_{280} for the unfolded protein is a good first approximation for the extinction coefficient of the native protein for many qualitative purposes. More quantitative approaches require determination of the exact extinction coefficient of the native protein. We can determine the extinction coefficient for the protein in its folded state by comparing the absorbance of a protein solution at the same concentration under denaturing and native conditions (Box 21.3). From the Lambert-Beer law (eq. 21.17), it is clear that the ratio of the absorption A_{native} and $A_{\text{denatured}}$ equals the ratio of the extinction coefficients $\varepsilon_{\text{native}}$ and $\varepsilon_{\text{denatured}}$ when path length and concentrations are identical.

BOX 21.3: DETERMINATION OF EXTINCTION COEFFICIENTS FOR NATIVE PROTEINS ACCORDING TO GILL & VON HIPPEL.

Gill & von Hippel developed a straightforward procedure for the determination of protein extinction coefficients under native conditions. The only information needed is the protein sequence (or at least the number of the chromophoric groups tyrosine, tryptophan, and cystines), and a measurement of the absorption of the protein under native and denaturing conditions. The extinction coefficient for the denatured protein can be calculated from the number of tryptophans (n_{Trp}), tyrosines (n_{Tyr}) and cystines ($n_{Cystine}$), and the extinction coefficients ε_{Trp}, ε_{Tyr}, and $\varepsilon_{Cystine}$:

$$\varepsilon_{denatured} = n_{Trp} \cdot \varepsilon_{Trp} + n_{Tyr} \cdot \varepsilon_{Tyr} + n_{Cystine} \cdot \varepsilon_{Cystine} \qquad \text{eq. 21.18}$$

Gill & von Hippel determined the extinction coefficients ε_{Trp}, ε_{Tyr} and $\varepsilon_{Cystine}$ in a 20 mM phosphate buffer, pH 6.5, 6 M guanidinium chloride, using N-acetyl-tryptophanamide as a reference substance for tryptophan, a Gly-Tyr-Gly peptide for tyrosine, and cystine for cysteine (Table 21.1).

TABLE 21.1:
Extinction coefficients for model compounds.

compound	$\varepsilon_{280\ nm}$ ($M^{-1}\ cm^{-1}$)
N-acetyl-tryptophanamide	5690
Gly-Tyr-Gly	1280
cystine	120

Values from Gill & von Hippel (1989) Anal. Biochem. 182(2): 319–326.

The absorption of the denatured protein, $A_{denatured}$, and of the native protein, A_{native}, can be described by the Lambert-Beer law as

$$A_{denatured} = \varepsilon_{denatured} \cdot c \cdot d \qquad \text{eq. 21.19}$$

and

$$A_{native} = \varepsilon_{native} \cdot c \cdot d \qquad \text{eq. 21.20}$$

For identical concentrations c and path lengths d, the extinction coefficient of the native protein can then be calculated from the absorbances under native and denaturing conditions, A_{native} and $A_{denatured}$, according to

$$\varepsilon_{native} = \frac{A_{native}}{A_{denatured}} \cdot \varepsilon_{denatured} \qquad \text{eq. 21.21}$$

This extinction coefficient ε_{native} of the native protein can then be used in concentration determinations of the protein under native conditions, without the need for prior denaturation. It is important to note that the calculation of the extinction coefficients relies on identical concentrations of native and denatured proteins. It is therefore advisable to repeat the absorption measurements with independently diluted samples.

Sometimes a protein of interest does not contain chromophoric amino acids. In this case, the photometric determination of protein concentrations following Ehresmann provides an alternative. This method measures the absorption of the protein backbone and is thus independent of the protein sequence. The relation

$$c = \frac{A_{228.5} - A_{234.5}}{3.14} \qquad \text{eq. 21.22}$$

was established empirically to determine protein concentrations in plant extracts in the presence of nucleic acids. $A_{228.5}$ and $A_{234.5}$ denote the absorption at 228.5 nm and 234.5 nm, respectively. The extinction coefficient of nucleic acids is identical at these wavelengths, and contributions from nucleic acids thus cancel. The factor 3.14 was determined empirically and converts the absorption difference into a concentration in mg mL^{-1} (which can be converted into molarity using the molecular mass of the protein). An advantage of the Ehresmann method is its high sensitivity because of the high extinction coefficient and the large number of peptide bonds, and the small amount of sample needed.

Due to the high extinction coefficients of nucleic acids, absorption is also a sensitive probe to determine the concentrations of DNA or RNA. Concentration determination of nucleic acids is hampered by the strong hypochromic effect upon duplex formation, and by the dependence of the extinction coefficient of the bases on their neighbors. For larger DNA with an approximately equal distribution of A, G, C, and T, one absorption unit at 260 nm (d = 1 cm) corresponds to approximately 55 µg mL^{-1} double-stranded DNA, or 33 µg mL^{-1} single-stranded DNA. For RNA, one absorption unit is equivalent to a concentration of approximately 40 µg mL^{-1}. If the sequence (or base composition) of the RNA is known, the concentration can be determined from the absorption of the nucleotide solution after alkaline hydrolysis of the RNA.

21.2.7.2 Spectroscopic assays for enzymatic activity

The facile conversion of absorption into concentrations also renders it a sensitive probe for measuring enzymatic activity when the natural substrate and the product of an enzymatic reaction differ in absorption. Examples are enzymes that catalyze redox reactions and transfer electrons to or from flavins or NAD(P)/NAD(P)H, such as oxidases and dehydrogenases. In other cases, *chromogenic substrates* can be used. Chromogenic substrates do not absorb, but enzymatic turnover generates a product that then shows absorption (Figure 21.16). Simple examples are the blue/white screening for expression of β-galactosidase in bacteria using the galactose-derivative 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal), and the ortho-nitrophenyl-β-galactoside (ONPG) test for β-galactosidase activity. Chromogenic peptides are frequently used as substrates to monitor peptidase or proteinase activity.

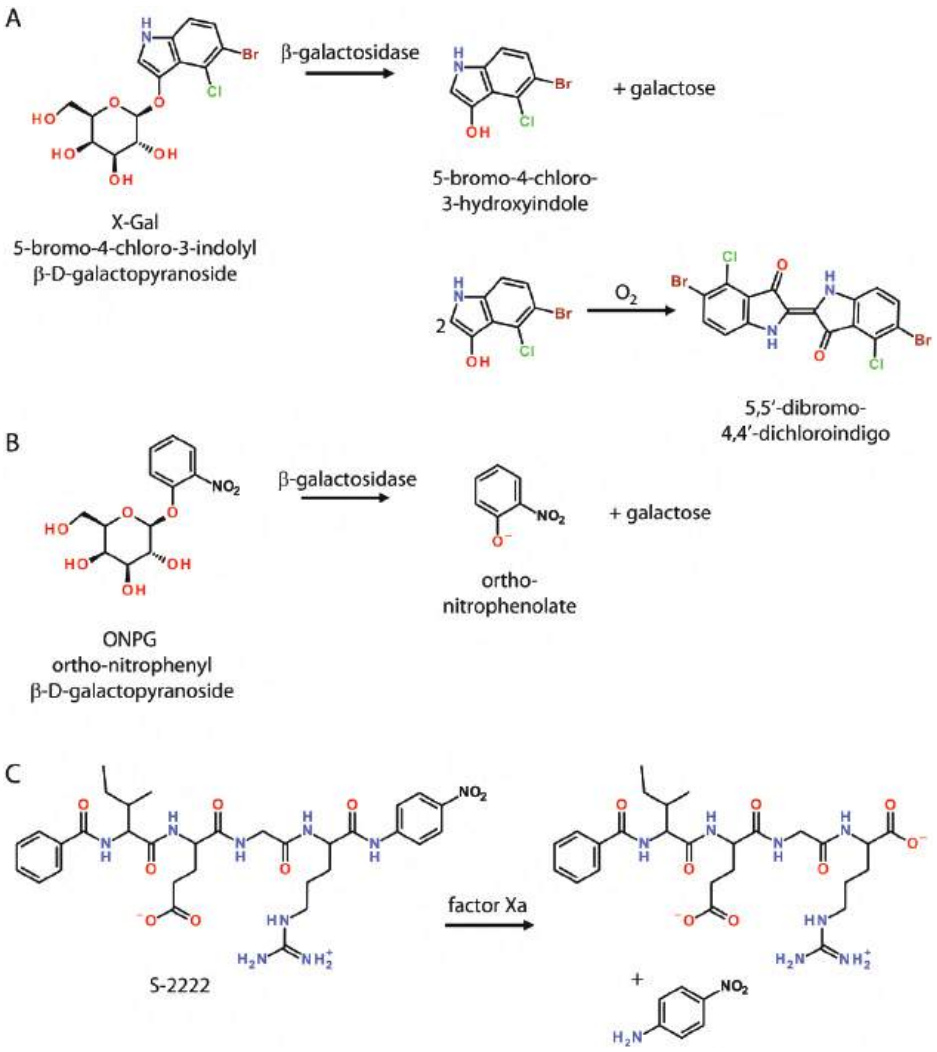


FIGURE 21.16: Chromogenic substrates and their reactions. A: 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-Gal) is a chromogenic substrate for the enzyme β-galactosidase. β-galactosidase cleaves the glycosidic bond in X-Gal, and generates 5-bromo-4-chloro-3-hydroxyindole. Two molecules of 5-bromo-4-chloro-3-hydroxyindole are oxidized to 5,5′-dibromo-4,4′-dichloroindigo, a blue insoluble pigment. B: ortho-nitrophenyl-β-D-galactopyranoside is cleaved to ortho-nitrophenolate and galactose by β-galactosidase. The reaction can be monitored by the strong absorption of ortho-nitrophenolate at 420 nm. C: The activity of peptidases and proteinases can be monitored using chromogenic peptides. S-2222 is a tetrapeptide of the sequence Ile-Glu-Gly-Arg, covalently linked to para-nitro-aminobenzene by a peptide bond. Factor Xa, a protease of the blood coagulation cascade, cleaves the terminal peptide bond. Factor Xa activity can be monitored *via* the absorption of para-nitro-aminobenzene at 380 nm.

If the enzyme of interest does not turn over substrates with an accompanying change in absorption, the catalyzed reaction can be coupled to other reactions that lead to a spectroscopically measurable output. A *coupled enzymatic assay* is frequently employed to monitor ATP hydrolysis by ATPases (Figure 21.17). The absorption spectra of ATP and ADP are identical, and the hydrolysis of ATP to ADP and inorganic phosphate can therefore not be monitored directly *via* a change in absorption. However, the reaction can be coupled to the oxidation of NADH to NAD^+. The nicotinamide ring of this cofactor absorbs at 340 nm in the reduced, but not in the oxidized form. The decrease in absorption at 340 nm therefore provides a sensitive spectroscopic probe for reactions that can be coupled enzymatically to the oxidation of NADH. To couple the hydrolysis of ATP to the oxidation of NADH, the ADP that is generated by the ATPase of interest is converted back to ATP by the enzyme pyruvate kinase that uses phosphoenol pyruvate as a phosphoryl group donor.

Phosphoenol pyruvate is converted to pyruvate, which is then reduced to lactate by lactate dehydrogenase, an enzyme that uses NADH as the electron donor. Over all three reactions, one ATP hydrolysis event is coupled to the oxidation of one NADH molecule. With the known extinction coefficient of NADH, the observed decrease in absorption at 340 nm over time can be converted into a change in concentration (of NADH and thus of ATP) over time. This change in ATP concentration directly represents the reaction velocity v (in M s^{-1}). In coupled enzymatic assays, it is essential to adjust enzyme concentrations and experimental conditions such that the overall reaction velocity is determined by the enzyme of interest, and not limited by the helper enzymes in the coupled reactions (see Chapter 10).

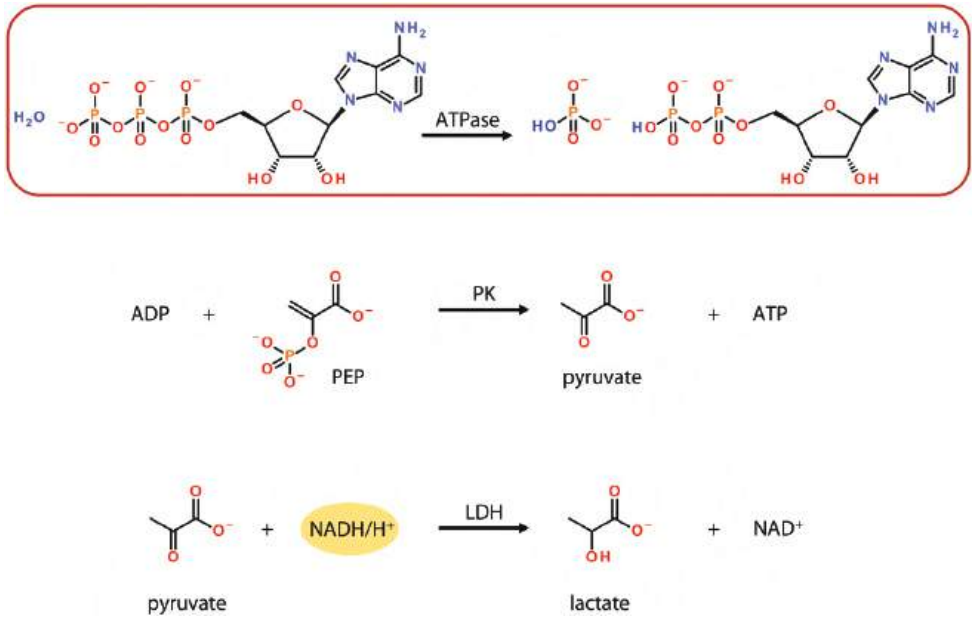


FIGURE 21.17: Coupled enzymatic ATPase assay. ATPases catalyze the hydrolysis of ATP to ADP and phosphate (top). The reaction is spectroscopically silent because ATP and ADP show the same absorption at 260 nm. In a coupled enzymatic assay, the hydrolysis of ATP is coupled to the oxidation of NADH: the ADP generated by ATP hydrolysis is phosphorylated to ATP by pyruvate kinase (PK) using phosphoenol pyruvate (PEP) as a phosphoryl group donor. The pyruvate generated by PK is then reduced to lactate by lactate dehydrogenase (LDH), with the concomitant oxidation of $NADH/H^+$ to NAD^+. NADH absorbs at 340 nm, but NAD^+ does not. Therefore, NADH consumption can be followed as a decrease in the absorption at 340 nm, and the concentration decrease can be calculated from the change in absorbance using the Lambert-Beer law. Per ATP hydrolyzed, one NADH is oxidized. The concentration decrease of NADH thus directly corresponds to the concentration decrease of ATP.

21.2.7.3 Spectroscopic tests for functional groups

Chromogenic reagents can also be used to test for the presence and reactivity of functional groups. One example is the counting of cysteines in proteins using the disulfide 5,5'-dithiobis-2-nitrobenzoic acid (DTNB, *Ellman's reagent*; Figure 21.18). DTNB forms mixed disulfides with solvent-accessible cysteines, which is associated with the release of the TNB^- ion which absorbs light at 412 nm. Per disulfide formed, one molecule of TNB^- is released. The molar amount of TNB^- that is generated can be calculated according to the Lambert-Beer law (eq. 21.17) from the increase in absorption upon incubation of the protein of interest with DTNB. The final concentration of TNB^- directly corresponds to the concentration of reactive cysteines. With the protein concentration known, the ratio of the concentration of free cysteines to protein gives the number of solvent-accessible cysteines per protein. If the experiment is performed under denaturing conditions, the total number of cysteines is determined. The difference between the two numbers (total-accessible) gives the number of cysteines buried in the protein interior. Cysteines are commonly used for site-specific chemical modification of proteins. The reaction of cysteines with maleimides of fluorescent dyes or spin labels

is widely used to introduce labels for fluorescence (Section 21.5) or electron paramagnetic resonance (EPR) spectroscopy (Section 22.2). Solvent-accessible cysteines that are suitable for modification can be identified by their reactivity with DTNB. Comparison of the number of reactive cysteines in a DTNB test under reducing and oxidizing conditions, on the other hand, yields information on disulfide bonds in the protein of interest.

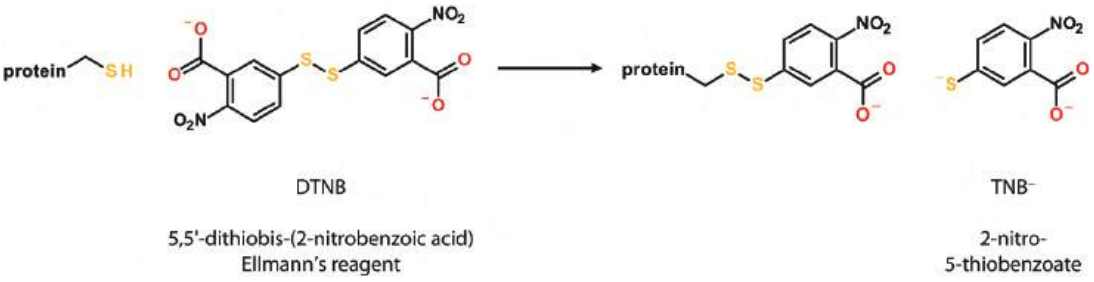


FIGURE 21.18: DTNB test for accessible cysteines. DTNB (5,5′-dithiobis-2-nitrobenzoic acid); shown is the anion), also called Ellman's reagent, forms mixed disulfides with cysteines. The 2-nitro-5-thiobenzoate (TNB^-) shows absorption at 412 nm. From the increase in absorption at 412 nm, the concentration of TNB^- can be calculated. Comparison with the protein concentration gives the number of free cysteines.

21.2.7.4 Absorption as a probe for structural changes

Absorption can also be used to monitor structural changes in proteins and nucleic acids because the absorption depends on the local environment of a chromophore (Section 21.2.4). The different absorption of folded and unfolded proteins can be exploited to follow protein folding and unfolding spectroscopically, and to determine the thermodynamic stability of proteins (Box 21.4).

BOX 21.4: ABSORPTION AS A SPECTROSCOPIC PROBE FOR PROTEIN UNFOLDING.

The absorption of aromatic amino acids depends on their chemical environment. In unfolded proteins, all aromatic amino acids are exposed to the aqueous solvent. Upon folding, at least some of these side chains become buried in the hydrophobic core of the protein. Therefore, the absorption of folded (native) and unfolded (denatured) protein differs. Absorption can thus be used as a spectroscopic probe to follow folding or unfolding of proteins and to determine the thermodynamic stability of proteins (Figure 21.19). An analysis of the equilibrium unfolding curves according to the two-state model of protein folding (Section 18.5.4, Figure 18.57) yields the thermodynamic parameters for folding. The general derivations of the relationships between the measured absorption A_{obs} and the thermodynamic parameters for protein folding for thermal and denaturant-induced unfolding are described in Section 18.5.4.

(*Continued*)

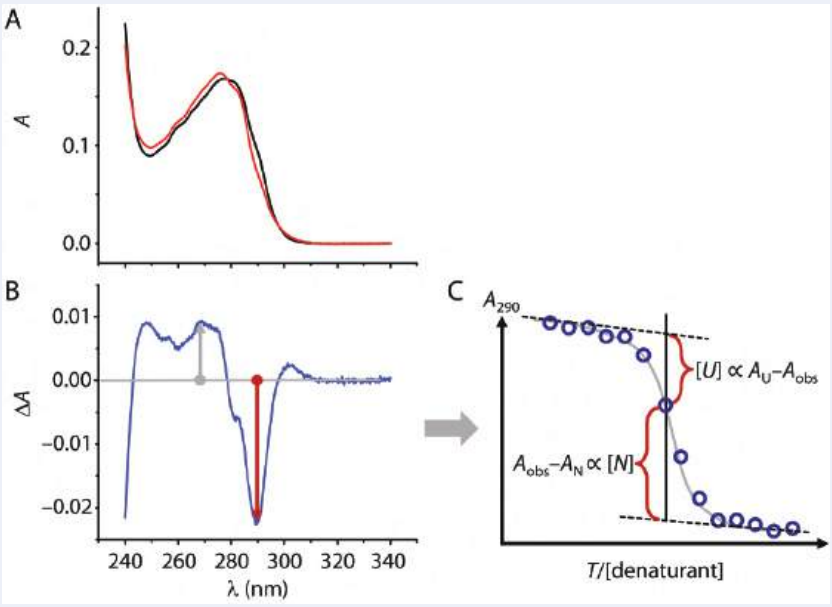


FIGURE 21.19: Protein unfolding monitored by absorption. A: The absorption of a native protein (black) differs from the absorption of the unfolded protein (red). Usually, a blueshift is observed upon unfolding. B: From the difference spectrum (denatured – native), the optimal wavelength for monitoring folding/unfolding can be determined (gray: maximum signal increase, red: maximum decrease). C: At 290 nm, the absorption decreases upon unfolding with increasing temperature (thermal denaturation) or denaturant concentration (chemical denaturation, Section 18.5 and Figure 18.53). For a two-state transition, the fraction of native and unfolded protein at each temperature or denaturant concentration can be determined from the difference of the measured absorption (A_{obs}) and the absorption of the native or unfolded protein (A_N, A_U; see Section 18.5.4 for the general derivation of a spectroscopic signal y_{obs} as a function of thermodynamic parameters for protein folding).

The strong hypochromic effect of DNA upon double helix formation provides a sensitive spectroscopic probe for DNA duplex stability. DNA melting curves, measured as the increase in the absorption at 260 nm as a function of temperature, define the melting temperature T_m at which 50% of the double-helical regions are disrupted (Box 21.5). A quantitative thermodynamic analysis requires taking into account the molecular mechanism of dissociation of the duplex into two single-strands.

BOX 21.5: ABSORPTION AS A PROBE FOR DNA DENATURATION.

Dissociation of a DNA duplex into two single strands is accompanied by an increase in absorption (hyperchromic effect). The increase in absorption serves as a spectroscopic probe for duplex dissociation, and can be used to compare melting temperatures and duplex stabilities (Figure 21.20).

(*Continued*)

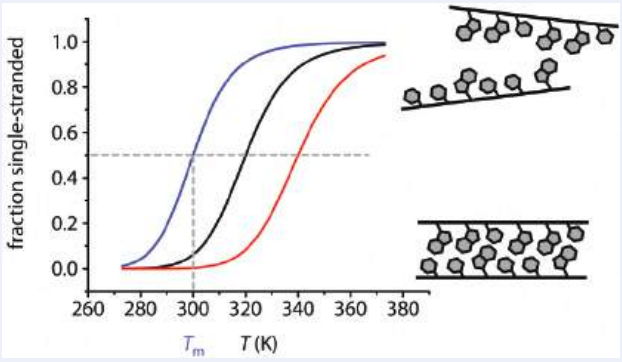


FIGURE 21.20: DNA melting curve. Melting curves for three DNA duplexes with increasing thermal stability (blue→black→red). With increasing temperature, the absorption of DNA increases because of dissociation of the duplex into the single-strands and the associated hypochromic effect. The inflection point where 50% of the DNA has dissociated is defined as the melting temperature T_m.

21.2.8 Potential pitfalls

Absorption measurements are in principle very straightforward to perform using standard absorption photometers and are easy to analyze. However, the dynamic range of photometers is typically limited to absorption values between 0.01 and 2. One reason for this limitation lies with the sensitivity of the detector: at very low absorption, a tiny difference between the intensities of the incident and the transmitted light has to be determined accurately. If the absorption is high, on the other hand, the little transmitted light will ultimately be below the detection limit. At $A = 3$ and $d = 1$ cm, only 0.1% of the incident light reaches the detector!

It is also important to note that deviations from the Lambert-Beer law with its linear relationship between concentration of the molecule of interest and the absorption can occur. The Lambert-Beer law only holds for ideal solutions of non-interacting molecules. In highly concentrated solutions, inter-molecular interactions may influence the charge distributions of neighboring molecules, altering their extinction coefficient. High concentrations of the absorber can also lead to changes in the refractive index of the solution and to deviations from the Lambert-Beer law. Chemical interactions, such as association/dissociation reactions, also may lead to non-linear relationships between concentration and absorption if the position of the equilibrium changes in the concentration range studied. Scattered light from larger particles in the sample, stray light, fluorescence, and phosphorescence can also cause deviations from linearity. Finally, the Lambert-Beer law is only valid for the absorption of monochromatic light, with a single, defined wavelength. If the excitation light used is not monochromatic, different extinction coefficients of the molecule of interest at the different wavelengths present in the excitation light, will also cause deviations from linearity. This effect can be minimized by exciting at the absorption maximum, where the wavelength dependence of the extinction coefficient is smallest.

Problems can also result from large particles in the sample, such as aggregates, that scatter the incident light (see Section 23.1). Scattering increases with the inverse of λ^4, and particularly distorts the characteristic protein spectrum in the lower-wavelength region (Figure 21.21). Apparent absorption above 300 nm is a strong indicator for the presence of scattering because aromatic residues in proteins do not show absorption in this spectral range.

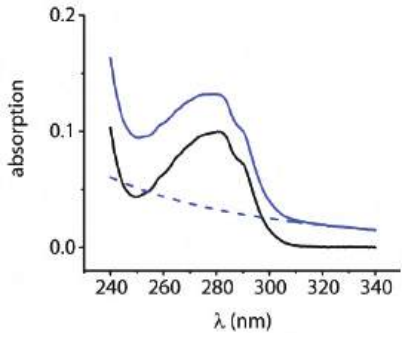


FIGURE 21.21: Scattering distorts absorption spectra. The intensity of scattered light (blue, broken line) increases with $1/\lambda^4$. When scattering particles are present, the measured spectrum of a protein (blue) will be distorted from the real protein absorption spectrum (black). Strong scattering not only changes the intensity of absorption, but also the apparent maximum.

21.3 LINEAR AND CIRCULAR DICHROISM

21.3.1 Linearly polarized light and linear dichroism

Absorption is based on an interaction of the dipole of the absorbing molecule with the electric field vector of the light. The strength of this interaction depends on the relative orientation of the electric field vector and the transition dipole of the molecule. In linearly polarized light, the oscillation of the electric field vector is confined to a plane, the *polarization plane* (Figure 21.22). Molecules oriented such that their transition dipoles are parallel to the polarization direction of the incident light will then preferentially absorb. The probability of a molecule to absorb polarized light depends on the magnitude of the vector component of the transition dipole that is parallel to the polarization direction, and thus on the angle θ between polarization direction and dipole moment (Figure 21.22). The absorption probability is proportional to $\cos^2\theta$ (Box 21.6). In solution, molecules can freely rotate, and have a random orientation of their transition dipoles. The number of molecules with an orientation between θ and dθ is proportional to $\sin\theta\, d\theta$. The overall *absorption probability* $f(\theta)$ then depends on the angle θ of the transition dipole and the polarization direction according to eq. 21.23:

$$f(\theta)d\theta = \cos^2\theta \sin\theta\, d\theta \qquad \text{eq. 21.23}$$

This angular dependence of the absorption probability leads to the preferential excitation of molecules whose transition dipoles have a large component parallel to the polarization of the incident light. In solution, only those molecules are selectively excited, a phenomenon called *photoselection* (Figure 21.22).



FIGURE 21.22: Photoselection. A: In linearly polarized light, the electric field vector is confined to the polarization plane (here: *xz*-plane). B: Left: The absorption probability depends on the angle θ between the transition dipole (blue arrow) and the polarization direction of the incident light (along the *z*-axis). Right: The angular dependence of the absorption probability is qualitatively represented by the intensity of the blue arrows that represent the transition dipoles. The orientations of dipole moments that lead to preferred absorption correspond to a dumbbell shape (gray). C: Chromophores in solution are randomly oriented (left). Their transition dipoles (arrows) are fixed with respect to the molecular geometry, and show random orientation. Upon excitation with light that is linearly polarized in the *z*-direction, those molecules whose transition dipole is oriented parallel to the *z*-axis absorb preferentially (red). As a consequence, the transition dipoles of the excited molecules are aligned. This process is called photoselection. Note that for simplicity excitation is shown only for molecules that are perfectly aligned with the polarization direction of the excitation light.

The dependence of the absorption on the polarization direction of the incident linearly polarized light is called *linear dichroism* (LD):

$$LD(\lambda) = A_v(\lambda) - A_h(\lambda) \qquad \text{eq. 21.24}$$

$A_v(\lambda)$ denotes the absorption of light polarized in the *z*-direction (vertical), and $A_h(\lambda)$ is the absorption of light polarized horizontally (along the *y*-axis). Often, the linear dichroism is converted to the *reduced linear dichroism* RD by normalization with the total absorption at this wavelength

$$RD(\lambda) = \frac{LD(\lambda)}{A(\lambda)} \qquad \text{eq. 21.25}$$

This expression is also called the *dichroic ratio*. Note that transition dipoles will be randomly oriented in solution (Figure 21.22). The total absorption over all molecules in solution will therefore not depend on the polarization direction of the incident light. Linear dichroism can only be observed for oriented molecules. The orientation of elongated molecules in solution can be achieved by applying flow. Polyelectrolytes or molecules with permanent electric dipoles, such as DNA double helices or helical peptides, can also be oriented in an electric field.

LD measurements are useful to examine whether an absorption band is caused by a single electronic transition between two states *a* and *b*, or if it is the result of two overlapping transitions, a→b and a→c. It is unlikely that the transition dipoles for these two transitions have identical directions. Therefore, the dichroic ratio is independent of the wavelength if only one transition is involved, but will depend on the wavelength for two overlapping transitions.

If the direction of the transition dipole is known, linear dichroism measurements can provide structural information. For example, the dipole of aromatic rings usually lies in the plane of the conjugated system. The orientation of aromatic rings within the structure can then be analyzed. An example is the orientation of nucleobases within a DNA duplex. In double-helical DNA, the nucleobases are oriented nearly perpendicular to the helical axis, and show strong absorption of light that is polarized in this direction, but no absorption of light that is polarized parallel to the helical axis. In addition, each base pair is rotated relative to the next, and shows different absorption of light polarized in different directions within the plane perpendicular to the helical axis (Figure 21.23). Helical peptides show linear dichroism in the infrared range (see Section 21.4.4).

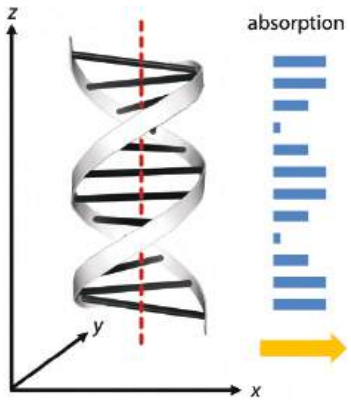


FIGURE 21.23: Linear dichroism of DNA. A DNA double helix with its helical axis (red) along the *z*-direction will only absorb light polarized in the *xy*-plane, the plane of the bases and their transition dipoles. Because of the rotation of each base pair by 360°/10.5 with respect to the neighboring base pair, the component of the transition dipole along the *x*-axis changes periodically. The resulting periodic pattern of absorption intensities for linearly polarized light with polarization along the *x*-axis is depicted on the right (blue bars).

Often, bases are tilted from their orientation exactly perpendicular to the helical axis, giving rise to a small component of the transition dipole parallel to the DNA helical axis (Figure 19.16). The inclination angle can be determined from LD measurements of DNA oriented in an electric field. Conversely, if structural information is available, LD measurements provide information on the direction of the transition dipole. This is possible in crystals with identical orientation of the chromophores, or with chromophores immobilized on surfaces or embedded in a matrix (Box 21.6).

BOX 21.6: DETERMINATION OF TRANSITION DIPOLE DIRECTIONS.

The probability of absorption is related to the square of the transition dipole induced in a molecule. Let us assume a molecule is oriented with its transition dipole at an angle θ with respect to the *z*-axis, and we excite this molecule with linearly polarized light along the *z*-axis. The dipole induced in the molecule is proportional to cosθ (Figure 21.24), and the resulting probability for a transition is proportional to $\cos^2\theta$. If we now rotate the polarization direction of the incident light in the *xz*-plane, the absorption of the molecule will be modulated by $\cos^2\theta$. From this modulation of absorption, the angle θ and thus the direction of the transition dipole can be determined. Using this approach, the relative orientations of chlorophyll molecules in a bacterial photosynthetic light-harvesting complex have been determined. Hofmann *et al.* (2003) Phys. Rev. Lett. 90: 013004.

(*Continued*)

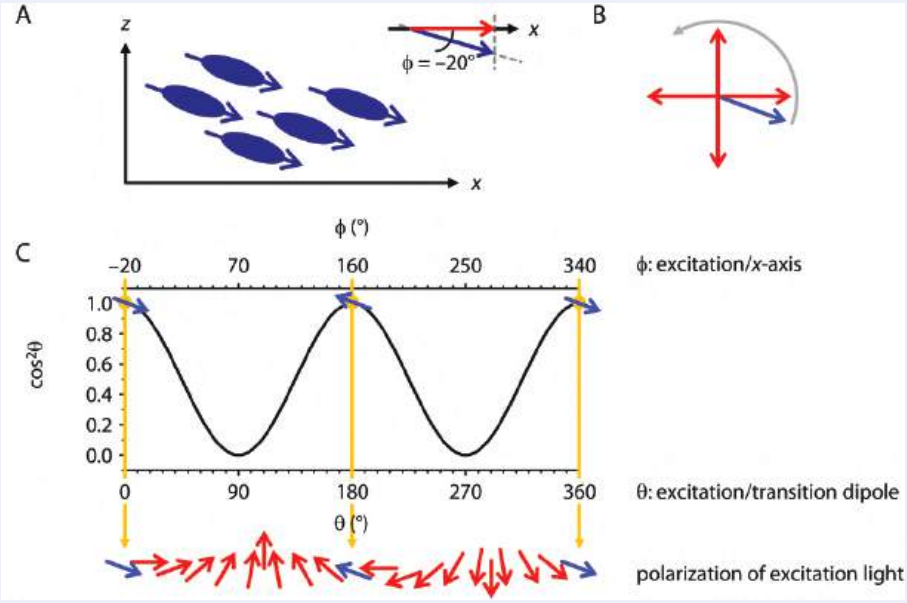


FIGURE 21.24: Determination of transition dipole directions. A: The transition dipole direction can be determined for aligned molecules (or for single molecules, see Section 25.7). B: The polarization direction of the excitation light is rotated, and the fluorescence emission from the molecules is measured for each polarization direction. C: The absorption and thus the fluorescence emission depends on the square of the cosine of the angle between the transition dipole and the polarization direction of the excitation light. From the modulation of the intensity with the direction of the polarization, the direction of the transition dipole can be determined.

21.3.2 Circularly polarized light and circular dichroism

Optical activity is a property that is linked to molecular asymmetry. The most common examples are asymmetric carbons in biomolecules, such as the Cα in amino acids and proteins (Section 18.1) or in sugars and nucleic acids. These asymmetric carbons have effects on chromophores in their vicinity. Optically active molecules alter the properties of light, leading to the phenomena of circular birefringence and optical rotation, or circular dichroism and ellipticity. Before we take a closer look at optical activity, we need to understand the concept of *circularly polarized light* and its relationship to linearly polarized light. Conceptually, we can regard linearly polarized light as the superposition of a *left-handed* and a *right-handed circularly polarized wave* of the same wavelength and phase (LHCPL, RHCPL; Figure 21.25). At any given point in time, the two vectors of the two circularly polarized waves will superimpose to a vector in the polarization plane of the resulting linearly polarized light.

In circularly polarized light, the tip of the electric field vector describes a helix in time and space (Figure 21.25), which is left-handed for left-handed circularly polarized light, and right-handed for right-handed circularly polarized light. This helical movement is the result of the superposition of the circular change in polarization direction and the propagation direction of the wave. Note that for linearly and circularly polarized light, the spatial repetition unit is the wavelength, and the repetition unit in time is the frequency. In linearly polarized light, the magnitude of the electric field vector is modulated, but its direction is constant. For circularly polarized light, the magnitude is constant, but the direction is modulated.

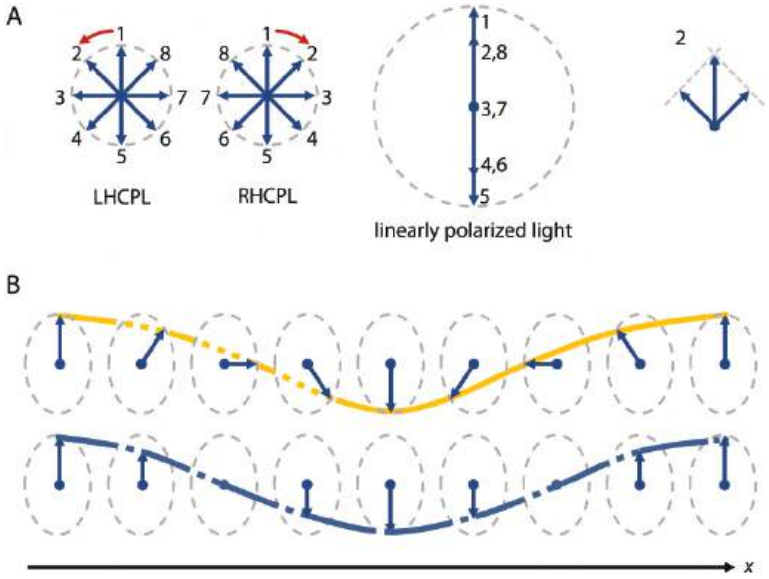


FIGURE 21.25: Circularly polarized light. A: In linearly polarized light, the oscillation of the electric field vector is confined to one direction, here in the vertical direction. At any given time (numbered from 1 to 8), the electric field vector can be generated by vector addition of two circularly polarized components of equal size, phase, and velocity, but opposite orientations. An example for time point 2 is shown on the right. One set of these vectors rotates in clockwise direction, the second set in counter-clockwise direction. The set of vectors whose tip rotates clockwise (right-handed) forms right-handed circularly polarized light (RHCPL), the other set forms left-handed circularly polarized light. Thus, linearly polarized light can be regarded as a superposition of RHCPL and LHCPL. B: In circularly polarized light, the electric field vector is modulated in direction, not in amplitude. Superimposed with the propagation in the *x*-direction, the tip of the vector describes a helix (yellow). In linearly polarized light, the direction of the electric field vector is fixed, but the amplitude is modulated. Superimposed with the propagation along the *x*-axis, its tip describes a cosine function (blue).

Experimentally, circularly polarized light can be generated by superposition of two linearly polarized waves of the same wavelength λ, but with a 90° angle between their polarization planes and a λ/4 phase shift. Depending on the direction of the phase shift (–λ/4 or +λ/4), superposition generates left- or right-handed circularly polarized light.

If molecules interact differently with the two circularly polarized components of linearly polarized light, the light transmitted by the sample will not remain linearly polarized. There are two possible effects. If the refractive indices for RHCPL and LHCPL are different ($n_L \neq n_R$), the two components propagate with different velocities while within the sample. After exiting the sample, RHCPL and LHCPL again travel at identical velocities. The light leaving the sample is still linearly polarized, but the polarization plane is rotated with respect to the polarization of the incident light by an angle ϕ (Figure 21.26). This phenomenon is called *circular birefringence* or *optical rotation dispersion* (ORD). The term *dispersion* takes account of the wavelength dependence of this effect. ORD is measured in degrees, *i.e.* the angle by which the polarization plane has been rotated by the sample. If RHCPL and LHCPL are absorbed differently ($A_L \neq A_R$ and $\varepsilon_L \neq \varepsilon_R$), the light leaving the sample is elliptically polarized: the tip of its vector traces an ellipse (Figure 21.26). This phenomenon is called *circular dichroism* (CD). CD is also measured in degrees (eq. 21.26). CD can only be observed at wavelengths where the sample absorbs ($\varepsilon_R \neq \varepsilon_L$). In contrast, ORD can be observed in regions outside absorption bands, and can be measured at high concentrations without inner-filter effects (Box 21.13). CD and ORD can be interconverted using the Kronig-Kramers transformation. We will only treat CD in the following because of its higher relevance for structural analysis of biomolecules.

The difference in absorption of RHCPL and LHCPL, Δε, that leads to circular dichroism is typically < 10 M^{-1} cm^{-1}, whereas ε of the respective molecule easily exceeds 20000 M^{-1} cm^{-1}. Because of the different absorption, the *E*-field vectors of RHCPL and LHCPL have different amplitudes when

they leave the sample. We can formally construct the elliptically polarized light leaving a sample that exhibits circular dichroism by adding these vectors at each time point (Figure 21.26). The ellipse that is described by the tip of the *E*-field vector is characterized by the larger and smaller half-axes, *a* and *b*. Their ratio *b*/*a* is called the *elliptic ratio*. CD is measured as ellipticity θ (in degrees), which is defined as the angle whose tangent is the elliptic ratio:

$$\theta = \arctan\left(\frac{b}{a}\right) \qquad \text{eq. 21.26}$$

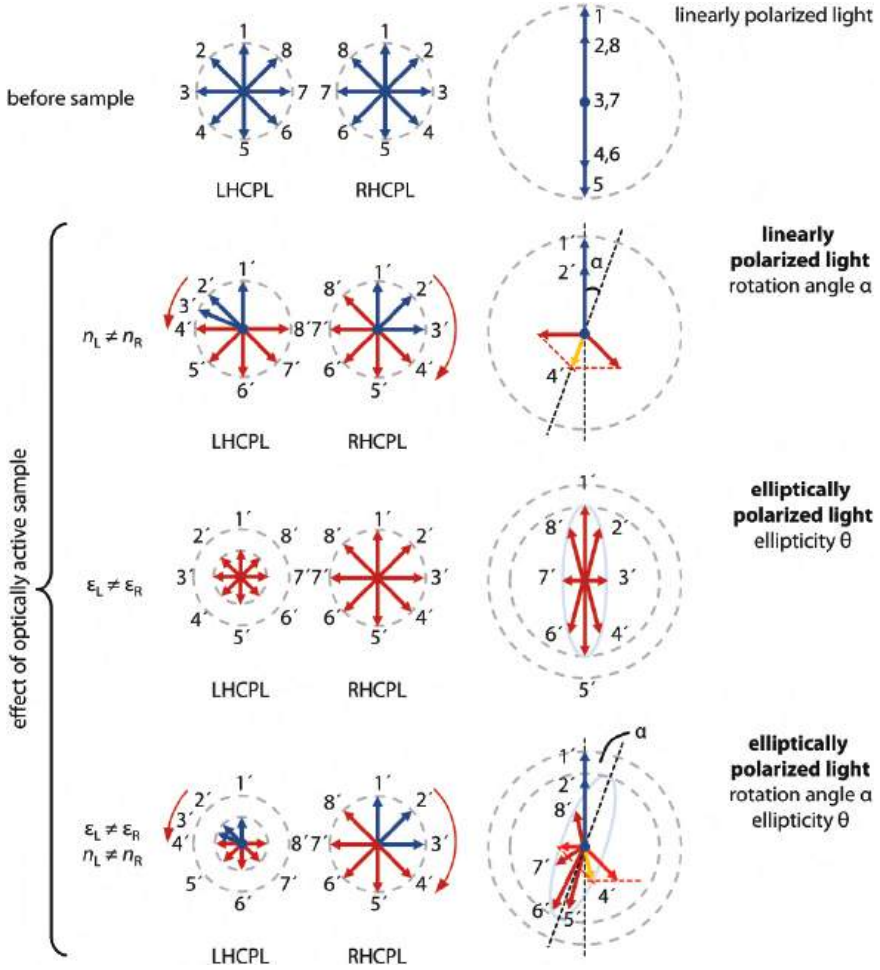


FIGURE 21.26: Circular dichroism and circular birefringence. An optically active sample is illuminated with linearly polarized light, composed of RHCPL and LHCPL (top). If the refractive indices for the RHCPL and LHCPL components are different ($n_L \neq n_R$), RHCPL and LHCPL travel with different velocities while within the sample (between time points 2 and 4; red, curved arrows), and have a constant shift after exiting the sample (time points 4′–8′, red). Superposition of the RHCPL and LHCPL components (illustrated for time point 4′, with the resulting vector depicted in orange) after the optically active sample gives linearly polarized light, but the polarization plane is rotated by the angle α (optical rotation dispersion, ORD). If the RHCPL and LHCPL components are absorbed to different extents ($\varepsilon_L \neq \varepsilon_R$; here, the RHCPL is not absorbed, but the LHCLP is), superposition of RHCPL and LHCPL after the sample gives elliptically polarized light. The ellipticity θ is defined by the ratio of the axes of the ellipse. If both $n_L \neq n_R$ and $\varepsilon_L \neq \varepsilon_R$, the light after the sample is elliptically polarized, and the plane of the ellipse is rotated by the angle α. Superposition of RHCPL and LHCPL after the sample is illustrated for time point 4′, with the resulting vector depicted in orange, and for time points 5′–8′ (only resulting vectors shown).

To compare ellipticities between different samples, the measured values are converted into the *molar ellipticity* θ_{molar} that is directly proportional to the difference of the extinction coefficients $\Delta\varepsilon$ for the RHCPL and LHCPL components:

$$[\theta]_{molar} = \frac{100\,\theta}{c \cdot d} \qquad \text{eq. 21.27}$$

c is the molar concentration, and d is the path length of the cuvette. The factor 100 is present for historical reasons and gives a unit for $[\theta]_{molar}$ of deg cm^2 dmol^{-1}. The ellipticities θ and θ_{molar} are proportional to the difference in extinction coefficient $\Delta\varepsilon$. For the molar ellipticity, the relationship is

$$[\theta]_{molar} \approx 3300\Delta\varepsilon \qquad \text{eq. 21.28}$$

(derivation in Box 21.7).

BOX 21.7: DERIVATION OF THE RELATIONSHIP BETWEEN THE ELLIPTICITY θ AND THE ABSORPTION DIFFERENCE ΔE.

We can express the half axes of the ellipse by the electric field vectors of LHCPL and RHCPL, E_L and E_R (Figure 21.26). The long axis a is given by $E_L + E_R$, the short axis b by $E_L - E_R$:

$$\theta = \arctan\frac{E_L - E_R}{E_L + E_R} \qquad \text{eq. 21.29}$$

Measured values for θ are very small, usually in the range of millidegrees, and so $\tan\theta$ can be approximated by θ. Furthermore, the intensities I_L and I_R are proportional to the square of the electric field vectors, and we can substitute

$$\theta = \frac{\sqrt{I_L} - \sqrt{I_R}}{\sqrt{I_L} + \sqrt{I_R}} \qquad \text{eq. 21.30}$$

According to the definition of the absorption (eq. 21.16 and eq. 21.17), we can express the intensities I_L and I_R in terms of A_L and A_R and I_0:

$$\log\frac{I_0}{I_L} = A_L = -\log\frac{I_L}{I_0} \qquad \text{eq. 21.31}$$

Conversion of the decadic logarithm to the natural logarithm gives

$$A_L = -\log\frac{I_L}{I_0} = -\frac{\ln\frac{I_L}{I_0}}{\ln 10} \qquad \text{eq. 21.32}$$

and

$$I_L = I_0\exp(-A_L \ln 10) \qquad \text{eq. 21.33}$$

Substituting these expressions for I_L and I_R into eq. 21.30 yields

$$\theta = \frac{I_0\exp\left(-\frac{A_L \ln 10}{2}\right) - I_0\exp\left(-\frac{A_R \ln 10}{2}\right)}{I_0\exp\left(-\frac{A_L \ln 10}{2}\right) + I_0\exp\left(-\frac{A_R \ln 10}{2}\right)} \qquad \text{eq. 21.34}$$

(*Continued*)

from which I_0 cancels. We can now expand the expression

$$\theta = \left(\frac{\exp\left(-\frac{A_L \ln 10}{2}\right) - \exp\left(-\frac{A_R \ln 10}{2}\right)}{\exp\left(-\frac{A_L \ln 10}{2}\right) + \exp\left(-\frac{A_R \ln 10}{2}\right)} \right) \cdot \left(\frac{\exp\left(\frac{A_L \ln 10}{2}\right)}{\exp\left(\frac{A_L \ln 10}{2}\right)} \right) \qquad \text{eq. 21.35}$$

and obtain

$$\theta = \frac{1 - \exp\left(-\frac{A_R \ln 10}{2} + \frac{A_L \ln 10}{2}\right)}{1 + \exp\left(-\frac{A_R \ln 10}{2} + \frac{A_L \ln 10}{2}\right)} = \frac{1 - \exp\left(\frac{\Delta A \ln 10}{2}\right)}{1 + \exp\left(\frac{\Delta A \ln 10}{2}\right)} \qquad \text{eq. 21.36}$$

with $\Delta A = A_L - A_R$. ΔA is typically small, and we can approximate the exponential function e^x as $1+x$:

$$\theta \approx \frac{1 - \left(1 + \frac{\Delta A \ln 10}{2}\right)}{1 + \left(1 + \frac{\Delta A \ln 10}{2}\right)} \qquad \text{eq. 21.37}$$

Because of the small values for ΔA, the term in the denominator can be approximated as 1, leading to

$$\theta \approx \frac{\frac{\Delta A \ln 10}{2}}{2} = \frac{\Delta A \ln 10}{4} \qquad \text{eq. 21.38}$$

The last step is the conversion of θ in rad to degrees, which gives the ellipticity θ as

$$\theta = \frac{\Delta A \ln 10}{4} \cdot \frac{360}{2\pi} \approx 33 \Delta A \qquad \text{eq. 21.39}$$

and the molar ellipticity $[\theta]_{molar}$ as

$$\theta_{molar} = \frac{100 \Delta A}{c \cdot d} \cdot \frac{\ln 10}{4} \cdot \frac{360}{2\pi} \approx 3300 \Delta \varepsilon \qquad \text{eq. 21.40}$$

Thus, ellipticity and molar ellipticity $[\theta]_{molar}$ are proportional to the difference in extinction coefficient $\Delta\varepsilon$, and therefore proportional to the concentration of the chromophore.

Circular dichroism is the result of an interaction of a chromophore with the electric and magnetic components of light. The absorption intensity of a chromophore depends on the dipole strength, related to the square of the electric transition dipole (Section 21.2.2). The intensity of a CD band is linked to the *rotational strength*, related to the product of the electric and the magnetic transition dipoles. Similar to the dipole strength, the rotational strength for a transition from initial state *i* to final state *f* can be determined from a measured CD spectrum by integration over the band.

21.3.3 Instrumentation

A standard CD spectrometer consists of a light source, a monochromator, and a polarizer to generate monochromatic, linearly polarized light (Figure 21.27). Because $\Delta\varepsilon$ is so small it is not possible

to measure the absorption of LHCPL and RHCPL separately and then calculate the small difference by subtraction. Instead, a piezo-elastic element or an electro-optic crystal is used to periodically modulate the excitation light, such that LHCPL and RHCPL components are produced in an alternating manner. The transmitted light gives rise to a photocurrent by the detector. The average transmitted light causes a direct current (DC) component, and the differential absorption of LHCPL and RHCPL by an optically active compound is detected as a modulation of the transmitted light, leading to a superimposed small alternating current (AC) signal. This small signal is accumulated over many modulation periods, so very small absorption differences can be determined. A phase-sensitive detector separates the AC and DC components. The ratio (AC/DC) then yields the absorption difference $\Delta A = A_L - A_R$.

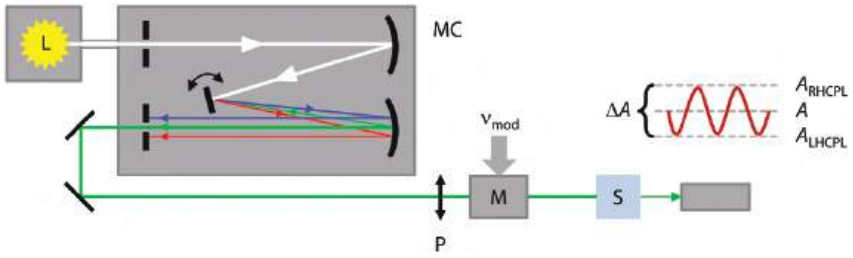


FIGURE 21.27: Principle of a CD spectrometer. The lamp (L) emits white light. The monochromator (MC) selects the desired excitation wavelength (green). After passing the polarizer (P), the light is linearly polarized. A modulator (M) generates an alternating excess of left- and right-handed circularly polarized light, alternating with the frequency ν_{mod}. Differential absorption of LHCPL and RHCPL then leads to a periodically modulated photocurrent on the detector. The DC component corresponds to the mean absorption *A*, the AC component reflects the difference in absorption ΔA of RHCPL and LHCPL.

21.3.4 Biological chromophores that show circular dichroism

Circular dichroism is observed with optically active, *i.e.* chiral molecules (Section 17.1). Asymmetric carbon atoms are abundant in biologically active molecules such as amino acids and nucleotides. Secondary structures formed by their respective polymers, proteins and nucleic acids, are also chiral. The most important chromophores in proteins are the peptide bonds and aromatic amino acid side chains. The peptide bond exhibits circular dichroism in the spectral region of its absorbance, which is the *far-UV* (190–250 nm; Figure 21.28). Although the peptide bond is not chiral on its own, it shows CD because it is embedded into an asymmetric environment in the secondary structure of folded proteins. The magnitude of this dichroism depends on the type of secondary structure element: each secondary structure has a characteristic CD spectrum. CD spectra from proteins of unknown structure can be described as a linear combination of such reference spectra to determine the fractions of secondary structure elements.

Aromatic amino acids in an asymmetric environment, *i.e.* in the interior of proteins where they are not free to rotate, show a CD signal in the *near-UV* spectral region (230–320 nm; Figure 21.28). In contrast to CD in the far-UV, near-UV CD is thus a useful probe for tertiary structure that immobilizes amino acid side chains. Nucleotides show a rather weak CD signal around 260 nm, which is typically positive for pyrimidine nucleotide and negative for purine nucleotides. The circular dichroism of DNA is caused by the periodic arrangement of the nucleobases in the double helix. Different helical forms, *e.g.* the A-forms and B-forms of DNA (see Section 19.5.1), give rise to different CD spectra. RNA exhibits CD characteristics similar to A-form DNA (Figure 21.28).

Similar to the near-UV CD of aromatic amino acid side chains, small non-optically active molecules can show induced CD upon binding to a macromolecule in an asymmetric environment. An example is the induced CD of pyridoxal phosphate when it is bound to aspartate aminotransferase.

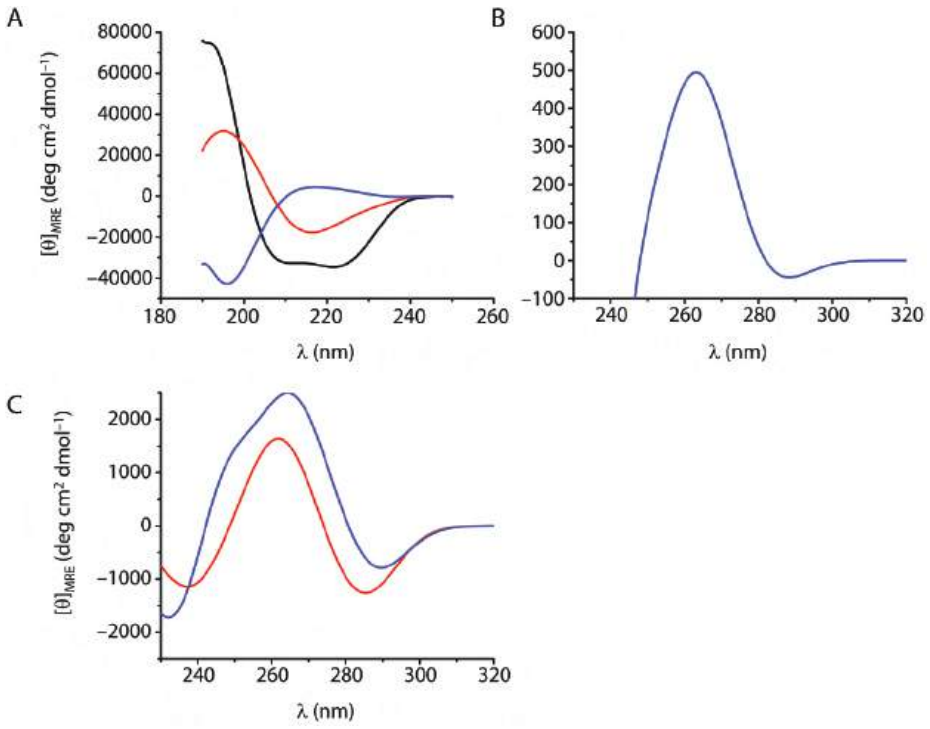


FIGURE 21.28: Circular dichroism of proteins and nucleic acids. A: Far-UV CD spectra of proteins. The CD spectrum of the peptide bond depends on the secondary structure. Black: α-helix, red: β-sheet, blue: random coil. Chou & Fasman (1974) Biochemistry 13(2): 211–222. B: Near-UV CD spectrum of a tryptophan-rich protein. Fixation of the tryptophan side chain in the asymmetric environment within the tertiary structure of the protein gives rise to circular dichroism. C: CD of DNA. The CD spectrum of DNA depends on the helix geometry (red: B-DNA, blue: A-DNA, see Section 19.5.1). Spectra in B and C are calculated using dichrocalc. Bulheller & Hirst (2009) Bioinformatics 25(4): 539–540.

21.3.5 Applications

CD spectra of the peptide bonds of folded proteins depend strongly on the secondary structure (Figure 21.28). Therefore, CD measurements are commonly used to determine the fraction of secondary structure elements in a protein of unknown structure. To this end, a CD spectrum of the protein of interest is measured, and described as the weighted sum of reference spectra. In the simplest case, the reference spectra are from model peptides that adopt a defined secondary structure (α-helix or β-sheet) or form a random coil (rc). The measured spectrum $x(\lambda)$ can then be described as

$$x(\lambda) = f_\alpha \cdot \alpha(\lambda) + f_\beta \cdot \beta(\lambda) + f_{rc} \cdot rc(\lambda) \qquad \text{eq. 21.41}$$

where f_α, f_β, and f_{rc} are the fractions of the protein that adopt α-helical or β-sheet structure or form a random coil, and $\alpha(\lambda)$, $\beta(\lambda)$, and $rc(\lambda)$ denote the corresponding reference spectra.

This rather simple procedure assumes that the entire structure of a protein can be represented by three types of secondary structure (α, β, or lack of regular structure, *i.e.* random coil), and that spectra of peptides appropriately reflect the CD properties of authentic proteins. However, the CD signal (per amino acid) of an α-helix depends on its length, and the CD spectrum of β-sheets is highly dependent on the environment. It differs for parallel and antiparallel β-sheets, and is affected by the commonly observed twist in β-sheets. Also, other secondary structure elements such as β-turns contribute to the CD spectrum of proteins. Finally, non-regularly structured elements in proteins are often not random coils, but have some degree of order. Altogether, the approach underlying eq. 21.41 is thus too simplistic. Modern programs therefore use a database of reference spectra that has been obtained from a principal component analysis of measured CD spectra of proteins with known structures.

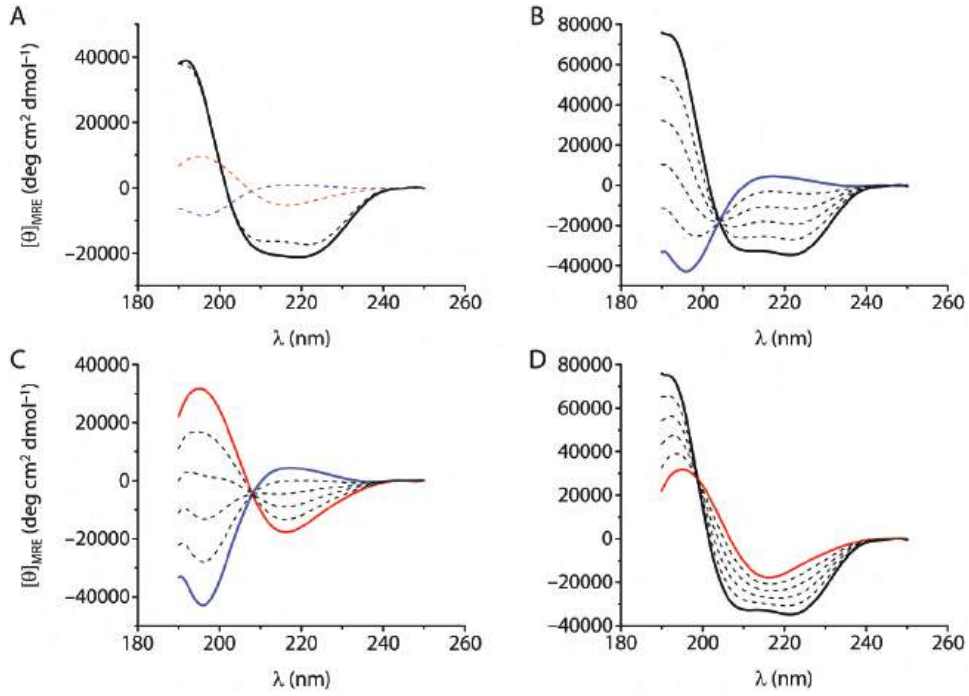


FIGURE 21.29: CD to analyze the secondary structure content of proteins. A: The CD spectrum of a protein of unknown secondary structure (black) can be described as a linear combination of the reference spectra (dotted lines) for α-helix (black), β-sheet (red) and random coil (blue). The reference spectra are multiplied by the fractions $f_\alpha = 0.5$, $f_\beta = 0.3$, and $f_{rc} = 0.2$. B: CD spectra for a helix→coil transition. The extreme spectra correspond to 100% α-helix (black) and 100% random coil (blue). The dotted spectra in-between reflect 80% α-helix/20% random coil, 60% α-helix/40% random coil, 40% α-helix/60% random coil, and 20% α-helix/80% random coil. C: CD spectra for a β-sheet→coil transition. The extreme spectra correspond to 100% β-sheet (red) and 100% random coil (blue). The dotted spectra in-between reflect 80% β-sheet/20% random coil, 60% β-sheet/40% random coil, 40% β-sheet/60% random coil, and 20% β-sheet/80% random coil. D: CD spectra for an α-helix→β-sheet transition. The extreme spectra correspond to 100% α-helix (black) and 100% β-sheet (red). The dotted spectra in-between reflect 80% α-helix/20% β-sheet, 60% α-helix/40% β-sheet, 40% α-helix/60% β-sheet, and 20% α-helix/80% β-sheet.

Circular dichroism is also a useful probe to follow protein folding or unfolding as a gain or loss in CD signal (Figure 21.29). Typically, the far-UV CD signal at 222 nm is used to follow the disappearance of α-helical structures upon protein unfolding. α-helix→β-sheet transitions can also be readily followed by CD (Figure 21.29). The near-UV CD signal (230–310 nm) originates from aromatic amino acid side chains in an asymmetric environment and is sensitive to changes in tertiary structure. CD has also been used to monitor structural changes of DNA and RNA (Figure 21.29), including thermal stability, metal ion binding, intercalation, and interactions with proteins.

21.3.6 Potential pitfalls

CD spectroscopy is a rather insensitive method that measures a small difference in absorption between LHCPL and RHCPL on top of a large average absorption. To increase the signal-to-noise ratio, spectra are often accumulated n-fold, which leads to an increase in signal by n, while the noise only increases by $\sqrt{n}$. Measurements of CD are based on accurate measurements of transmission, and the overall absorption should therefore be limited. The measured absorbance A_{obs}, and hence the signal-to-noise ratio, is maximal for $A = 0.8686$. Many common buffers including Tris show strong absorption near 200 nm, and buffers have to be selected carefully for far-UV measurements.

21.4 INFRARED SPECTROSCOPY

We have seen before (Section 21.2.1) that each electronic state comprises a set of vibronic states. Similar to electronic excitation, vibrations can be excited by radiation of the energy separating the two vibrational states involved. The energetic difference between two vibrational states is typically one order of magnitude smaller than the energy separating two electronic states. Therefore, vibronic excitation can be achieved with *infrared (IR) light*, at wavelengths of 1–10 μm. It is important to note that the energetic difference between the vibronic ground state and the first excited state is larger than the thermal energy, meaning IR spectroscopy at room temperature starts from V_0. While Raman spectroscopy probes vibrations indirectly as scattering (Section 23.1.3), IR spectroscopy directly measures transitions between vibrational energy levels of molecules due to absorption.

21.4.1 Bond vibrations: the harmonic oscillator

A chemical bond can be described as a *harmonic oscillator* (Figure 21.30; Section 17.2.1). We imagine a sphere attached to the ceiling by a spring. If we drop the sphere, it will fall until it has extended the spring, and will then be pulled back up again as the spring re-contracts. The force F on the sphere depends on the *stiffness* k of the spring and the current position x of the sphere according to *Hooke's law*:

$$F = -kx \qquad \text{eq. 21.42}$$

This force causes an acceleration of the particle that is

$$F = ma \qquad \text{eq. 21.43}$$

The acceleration is the derivative of the velocity v, which in turn is the derivative of the position x:

$$a = \frac{\mathrm{d}v}{\mathrm{d}t} = \frac{\mathrm{d}^2x}{\mathrm{d}t^2} \qquad \text{eq. 21.44}$$

By substituting eq. 21.44 into eq. 21.42 and then equating eq. 21.42 with eq. 21.43, we obtain

$$-kx = m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} \qquad \text{eq. 21.45}$$

Eq. 21.45 contains x and its second derivative, $\mathrm{d}^2x/\mathrm{d}t^2$. A possible solution for x is a sine function

$$x = b \sin \omega t \qquad \text{eq. 21.46}$$

with

$$\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -b\omega^2 \sin \omega t = -\omega^2 x \qquad \text{eq. 21.47}$$

We can substitute this expression into eq. 21.45, and obtain

$$-kx = m\left(-\omega^2 x\right) \qquad \text{eq. 21.48}$$

and the frequency of the oscillation as

$$\omega = \sqrt{\frac{k}{m}} \qquad \text{eq. 21.49}$$

or

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \qquad \text{eq. 21.50}$$

The potential energy of the sphere is then

$$E = \int F\,dx = k\int x\,dx = \frac{k}{2}x^2 \qquad \text{eq. 21.51}$$

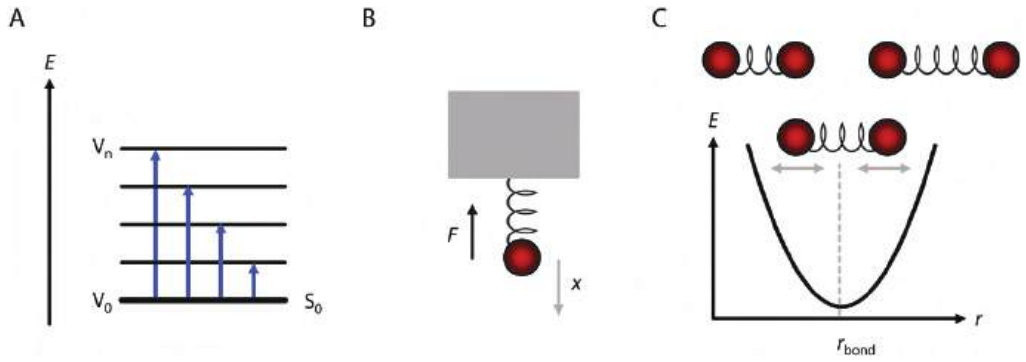


FIGURE 21.30: Vibration spectroscopy. A: Absorption of infrared light leads to transitions between vibrational states of a molecule in the electronic ground state S_0 (blue arrows). B: Oscillations of chemical bonds around their mean bond length r_{bond} can be described as harmonic oscillators. The potential energy of the harmonic oscillator increases parabolically with increasing deviations from the bond length (eq. 21.51, eq. 21.52).

We can now transfer the concept of a harmonic oscillator to the oscillation of a chemical bond (Section 17.2.1). Bond vibration can be described quantitatively by restricting the considerations to the area of the Lennard-Jones potential close to the distance where the potential energy is minimal. In this region, the potential can be approximated by a parabola (Figure 21.30), and we can express the potential energy E_{pot} as

$$E_{pot} = \frac{1}{2}k\left(r - r_{bond}\right)^2 \qquad \text{eq. 21.52}$$

with the force constant k, the actual distance r, and the bond length r_{bond}. Note that the term $(r\text{-}r_{bond})$ corresponds to the displacement x in eq. 21.42. The force constant k is a measure of the steepness of the walls of the potential, and the stiffness of the bond. The force during bond vibration follows Hooke's law:

$$F = k\left(r - r_{bond}\right) \qquad \text{eq. 21.53}$$

The *resonance frequency* ω of the harmonic oscillation of the bond depends on the reduced mass μ of the two atoms as

$$\omega = \sqrt{\frac{k}{\mu}} \qquad \text{eq. 21.54}$$

or

$$\nu_0 = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \text{eq. 21.55}$$

The reduced mass μ of the oscillator is defined as

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \qquad \text{eq. 21.56}$$

with the masses m_1 and m_2 of the two atoms. We can infer that bonds between atoms with larger masses absorb at lower frequencies. As an example, the IR absorption of a C-D bond appears at a lower frequency than the vibration of a C-H bond (isotope effect; see Chapter 10, Box 10.2). Stronger bonds with higher stiffness k show IR absorption at higher frequencies. Triple carbon-carbon bonds therefore absorb at higher frequencies than double or single bonds.

The energy of a harmonic oscillator is quantized, and the possible vibrational energy levels E_n depend on the vibrational quantum number n (1, 2, …):

$$E_n = h\nu_0\left(n + \frac{1}{2}\right) \qquad \text{eq. 21.57}$$

According to eq. 21.57, neighboring vibration levels are separated by

$$\Delta E = h\nu_0 \qquad \text{eq. 21.58}$$

and infrared absorption occurs when the frequency of the incident radiation equals the resonance frequency ν_0 of the bond. IR absorption is usually not given in frequency or wavelength, but in wavenumbers. The wavenumber $\tilde{\nu}$ is the inverse of the wavelength λ

$$\tilde{\nu} = \frac{1}{\lambda} = \frac{\nu}{c} \qquad \text{eq. 21.59}$$

and its dimension is typically cm^{-1}: it shows how many periodic waves occur over one centimeter. For example, the three water vibrations are at 3685 cm^{-1}, 3605 cm^{-1}, and 1885 cm^{-1}.

The advantage of wavenumbers lies with their direct proportionality to the energy of the radiation instead of the inverse proportionality between energy and wavelength (eq. 21.60)

$$E = h\frac{c}{\lambda} = hc\tilde{\nu} \qquad \text{eq. 21.60}$$

with the Planck constant h, and the speed of light c.

21.4.2 Molecule geometry, degrees of freedom, and vibrational modes

Vibration modes are only *IR-active*, *i.e.* give rise to an IR absorption band, when they are associated with a transition dipole moment ($\Delta\mu \neq 0$). Vibrations that are not associated with a transition dipole ($\Delta\mu \neq 0$) are *IR-inactive*. A vibration leads to a transition dipole if the permanent dipole of the molecule changes during the vibration. The number of possible vibration modes depends on the size and shape of the molecule. An atom has three *degrees of freedom*, corresponding to its translational movement along the three axes of a Cartesian coordinate system. A molecule that consists of N atoms has $3N$ degrees of freedom. Three of these degrees of freedom correspond to the translational movement of the entire molecule (*i.e.* of all atoms simultaneously) in the three dimensions. Molecules can also rotate: for a linear molecule, two degrees of freedom are assigned to rotation (around its long and its short axis), whereas a non-linear molecule can rotate around three different axes, corresponding to three degrees of freedom. Altogether, a linear molecule of N atoms has $3N–5$ degrees of freedom that are correlated with different types of vibrations (modes), and a non-linear molecule of N atoms has $3N–6$ different modes of vibration. Although the degrees of freedom increase tremendously with the size of the molecule, infrared spectroscopy is simplified by

the fact that functional groups show distinct vibration patterns and can be assigned even in spectra of rather complex molecules. We will therefore illustrate the different vibration modes with simple examples. First, we consider a diatomic molecule, such as H_2. The H_2 molecule can vibrate such that the two hydrogen atoms periodically come closer and move further apart. Such a vibration that leads to oscillations of the bond length is called stretching, and this mode of vibration is a *stretching mode*. It is the only vibration mode possible for this molecule ($3N–5 = 1$). H_2 does not show IR absorption, however, because its only vibrational mode is not associated with changes in the permanent dipole.

Now we consider two different molecules of three atoms, the linear CO_2 molecule ($3N–5 = 4$ vibrational modes), and the non-linear H_2O molecule ($3N–6 = 3$ vibrational modes; Figure 21.31).

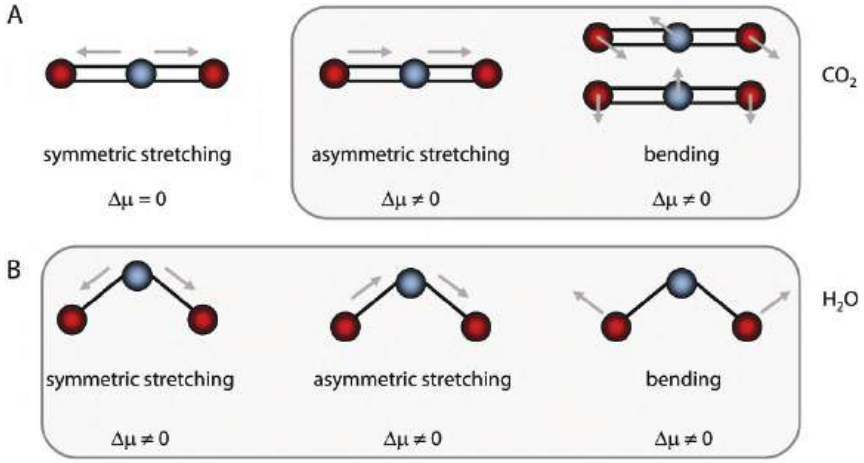


FIGURE 21.31: Vibrational modes of CO_2 and H_2O. A: CO_2 as a linear molecule has four vibrational modes. The four possible modes are the symmetric and asymmetric stretching vibrations, and two bending vibrations, one with the atoms moving into and out of the plane of the diagram, and a second with the atoms moving up and down. The symmetric stretching does not alter the dipole moment ($\Delta\mu = 0$) of the molecule and is thus IR-inactive. B: H_2O is a non-linear molecule with three vibrational modes. It undergoes symmetric and asymmetric as well as a bending vibration. All three modes lead to a change in the dipole moment ($\Delta\mu \neq 0$), and are thus IR-active.

In CO_2, the two bonds between the carbon and oxygen atoms can vibrate such that the two oxygen atoms move away and towards the central carbon in synchrony (*symmetric stretching* mode; Figure 21.31) or anti-correlated, with one oxygen moving toward the carbon while the second moves away from the carbon (*asymmetric stretching* mode). Symmetric stretching does not change the dipole moment of the molecule, and thus does not lead to IR absorption: it is IR-inactive. In contrast, the asymmetric stretching is accompanied by a change in the dipole moment of the CO_2 molecule, and this vibration gives rise to an IR band: it is IR-active. In addition to the stretching modes, CO_2 can undergo *bending* vibrations, where the two oxygen atoms are displaced into the same direction, such that the molecule is deformed from its linear shape, and the bending angle oscillates. A deviation from linearity alters the dipole of the molecule, and the bending vibration is therefore IR-active. CO_2 shows two possible modes of bending vibrations, with a relative displacement of carbon and oxygen in two different planes. These modes are isoenergetic and lead to one band in the IR spectrum with twice the intensity compared to the other asymmetric stretching mode.

In the non-linear water molecule, the three possible vibration modes are the symmetric stretching vibration, the asymmetric stretching vibration, and the bending vibration (Figure 21.31). All three vibration modes are associated with a change in the dipole moment of the H_2O molecule. Therefore, all three vibrations are IR-active, and give rise to absorption at 3685 cm^{-1} (symmetric stretching), 3605 cm^{-1} (asymmetric stretching), and 1885 cm^{-1} (bending).

Because of the requirement for changes in the molecular dipole for infrared absorption, highly polarized bonds such as CO-, CN-, PO-, and NH-bonds give rise to intense IR bands. The

resonances from vibrations of these bonds are found in the spectral region above 1500 cm^{-1}, and can be assigned to the individual bonds. Below wavenumbers of 1500 cm^{-1}, a large number of resonances corresponding to vibrations of the entire molecule are observed. This region is therefore called the *fingerprint region*. The fingerprint region is complex and very sensitive to the chemical identity of the molecule. Comparison of the fingerprint region with reference spectra can therefore be a very powerful method to identify molecules.

21.4.3 Instrumentation

An IR spectrometer consists of a lamp that emits radiation in the infrared spectral range, monochromators for selection of excitation and emission wavelength, and a detector. Quartz-tungsten-halogen lamps cover the spectrum up to wavelengths of 5 μm (2000 cm^{-1}), whereas heated resistive ceramic wires emit in the range from 1 μm up to several hundred micrometers (from 9500 cm^{-1} to tens of cm^{-1}). The wavelength is scanned by moving the diffraction grating of a monochromator.

IR spectra are typically depicted as % transmission as a function of the wavenumber of the incident light. IR absorption can be measured on solid, liquid, or gaseous material. Solids are usually mixed with KBr and compressed into a tablet. Liquids are spread between two NaCl plates, or are dissolved and measured in solution in NaCl cells. NaCl is transparent for infrared light. Special cells sealed with NaCl plates are used for measurements in the gas phase. Spectra are typically measured by ratio recording (Figure 21.32), where the transmission of the sample is measured relative to the transmission of a reference cell. The *n*-fold accumulation of spectra increases the signal-to-noise ratio.

Modern IR spectrometers excite all frequencies simultaneously and measure an *interferogram* from which the spectrum is obtained by *Fourier transformation* (FT; Section 31.15). The heart of an FT-IR instrument that enables simultaneous measurement of IR absorption at all frequencies is the *Michelson interferometer* that takes the place of the diffraction grating (Figure 21.32). In an FT-IR spectrometer, the excitation light first passes an interference plate. This plate functions as a beam splitter, and reflects half of the light to a fixed mirror, while the other half is transmitted toward a moveable mirror. The reflected light from both mirrors is superimposed on the interference plate, and then passes sample and reference. The light intensity as a function of the position of the moveable mirror, the *interferogram*, is detected by the detectors. To extract the spectrum, *i.e.* the intensity of the transmitted light as a function of the wavenumber, the interferogram has to be Fourier transformed from intensity as a function of distance into intensity as a function of inverse distance (wavenumbers). In NMR (Section 22.1.5), Fourier transformation is used to convert a signal as a function of time into a signal as a function of frequency (inverse of time), in X-ray crystallography (Chapter 24), it is employed to convert intensities in reciprocal space (inverse distance) to electron densities in real space (distance). With FT-IR, IR spectra are measured faster (within seconds), with higher signal-to-noise ratio, and with high precision in wavenumbers. Even faster processes can be followed by *step-scan FT-IR*. In this type of experiment, the reaction is triggered in the IR sample cell, for example by laser flash photolysis of a caged compound (see Section 28.3), and the time dependence of the change in IR absorption is measured at a fixed position of the moveable mirror in the Michaelson interferometer. The mirror is moved by a small step, and a new reaction is started to monitor the time dependence of the IR signal at this position of the mirror. This cycle is repeated until the entire interferogram has been sampled step by step. Individual time-dependent measurements are then combined to obtain a time-dependent IR spectrum with a time-resolution of as low as μs–ns, limited by the time needed to start the reaction and by the response time of the detector.

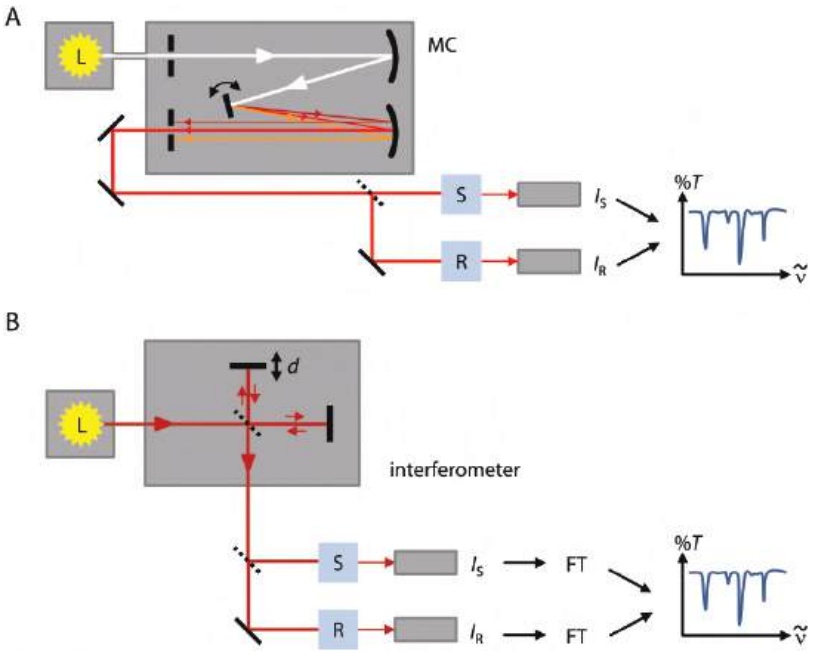


FIGURE 21.32: IR spectrometers. A: Ratio-recording IR spectrometer. The excitation wavelength is selected by passing the light from the light source (L) through a monochromator (MC) before irradiating sample (S) and reference (R) cell. The intensity of the transmitted light is measured as a function of the wavelength of the incident light. B: FT-IR spectrometer. The light from the light source (L) enters an interferometer. An interference plate reflects half of the light toward a moveable mirror (double-headed arrow), the other half passes toward a fixed mirror. The light reflected from both mirrors is recombined at the interference plate. Depending on the distance *d* the mirror has moved, the two wave trains have traveled different distances, and are phase-shifted relative to each other. Their superposition creates an interferogram. The light exiting the interferometer is then passed through sample (S) and reference cell (R). The absorbance of the sample alters the shape of the interferogram of the transmitted light. From the transmission as a function of the movement *d* of the mirror, the spectrum (transmission as a function of wavenumbers) can be calculated by Fourier transformation (FT).

21.4.4 Applications

To monitor changes in vibrations for individual bonds by IR spectroscopy, a particular resonance has to be assigned to a specific group. IR spectra of proteins are very crowded, but assignment of bands can be achieved by site-directed mutagenesis: when the amino acid whose side chain gives rise to a particular resonance is replaced by a different type of amino acid, the corresponding resonance will disappear, enabling assignment. The effect of isotope substitution on the vibration frequency can also be exploited for assignment. In studies of the mechanism of GTP hydrolysis by the oncoprotein Ras, GTP with ^{18}O substitutions at the α-, β-, or γ-phosphate has been used to assign phosphate bands to individual phosphates. Isotope labeling of proteins is achieved by supplying minimal media for recombinant protein production with isotopically labeled amino acids. Using this approach, isotope-labeled tyrosines have been incorporated into photoactive yellow protein, a bacterial blue light sensor, to record isotope-edited IR spectra.

Ligation by metal ions strongly polarizes bonds, and leads to an increase in their IR absorption. The interaction of CO, NO, or O_2 with the central iron in heme, for example, strongly affects their vibration frequencies, enabling studies on ligand binding to heme by IR spectroscopy.

The IR spectrum of proteins can be used to quantify secondary structure elements: the carbonyl CO stretching vibration of the peptide bond depends on the hydrogen bonding pattern and thus on the type of secondary structure element it is part of. The CO resonance of a protein can be described by a linear combination of the absorption bands for the CO stretching vibration in α-helices or β-sheets to quantify their fractions, similar to the secondary structure determination from far-UV

CD spectra (Section 21.3.5). While CD is very sensitive for α-helices, IR spectroscopy provides stronger signals for β-sheets, thus these methods are complementary.

The stretch vibration of backbone carbonyl groups (1652 cm^{-1}) in the peptide bond exhibits a transition dipole in the direction of the CO double bond. In helical peptides, this dipole is parallel to the helical axis, and helical peptides therefore show linear dichroism (Section 21.3.1) in the infrared, with a higher absorbance for infrared light polarized parallel to the helical axis. In contrast, the NH bending vibration is mixed with the CN stretch vibration (1549 cm^{-1}), leading to a transition dipole almost perpendicular to the helical axis. Here, a stronger absorption of light polarized perpendicular to the helical axis is observed.

IR spectroscopy can also be used to characterize cold-trapped reaction intermediates (cryo-FT-IR), for example in protein folding. Time-dependent changes in IR spectra are often difficult to interpret because of the large number of signals. Difference spectroscopy simplifies spectra significantly and allows one to focus on those signals that change in intensity upon conversion of one species into another. Maxima in a difference spectrum (product minus reactant) corresponds to a signal that is built up during the reaction, whereas minima (negative peaks) reflect signals that are present in the reactant but not in the product of the reaction. Step-scan FT-IR measures time-dependent changes in IR signals over the entire spectrum and has been used to identify structural changes in bacteriorhodopsin during the photocycle, and to dissect the phosphoryl transfer mechanism during GTP hydrolysis by the small GTP-binding protein Ras in the presence of its GTPase activating protein.

21.5 FLUORESCENCE

21.5.1 General considerations

Fluorescence is emitted when a molecule returns from the excited state to the ground state. While absorption starts from the ground state of a molecule, and therefore reports on molecular properties in the ground state, fluorescence is sensitive for all processes that occur while the molecule is in the excited state. Absorption and emission spectroscopy therefore yield complementary information.

Fluorescence is generally independent of the excitation wavelength. The reason for this lies in the time scale of the processes involved. We have seen before that absorption is a very fast process that occurs on the order of 10^{-15} s (Section 21.2.2). Depending on the energy of the excitation light, the molecule is excited from the singlet ground state S_0 to the S_1 state (or the higher electronic states S_2, etc.; Figure 21.33), and to different vibronic sub-states of S_1. Higher electronically excited states will relax to S_1 within 10^{-12} s by *internal conversion* (IC). On the same time scale, higher vibronic states of S_1 will relax to the V_0 level by *vibronic relaxation* (VR). The S_1,V_0 state is very long-lived (1–10 ns), and the fluorescence that occurs from this state is therefore independent of the wavelength of the excitation light. Molecules or moieties that exhibit fluorescence are termed *fluorophores*. Alternatively, molecules can relax to the ground state in non-radiative transitions. Molecules can also enter the triplet state T_1 from S_1 by *inter-system crossing* (ISC). In this transition, the electron spin is inverted; the transition is spin-forbidden. From T_1, the molecule returns to S_0 in a second spin-forbidden transition by emitting *phosphorescence*. Transitions between energy levels of fluorophores are often depicted schematically in a *Jablonski diagram* (Figure 21.33).

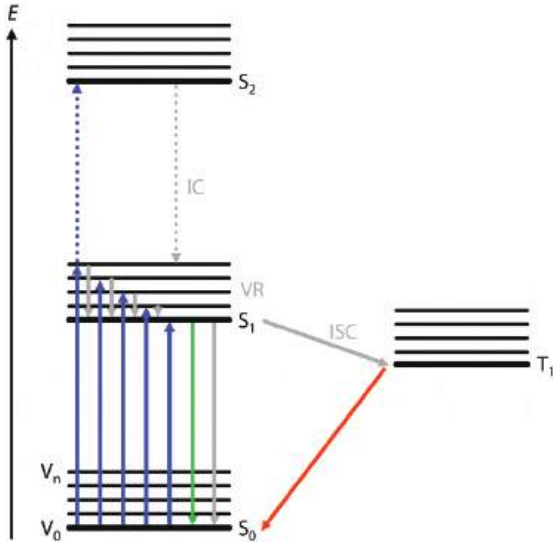


FIGURE 21.33: Jablonski diagram for fluorescence. Excitation from S_0,V_0 (blue arrows) leads to the population of different vibronic states V_0-V_n of the electronically excited S_1 state. Higher vibronic states of S_1 relax to the vibrational ground state S_1,V_0 by vibronic relaxation (VR, gray). Irradiation with light of higher energy (blue dotted line) can lead to the excitation of higher energy electronic states (S_2), followed by non-radiative relaxation to S_1 (internal conversion, IC; gray dotted arrow). Molecules return to the ground state by emitting fluorescence (green) or by non-radiative transitions (gray), or enter the triplet state T_1 from S_1 by inter-system crossing (ISC) and return to S_0 by emitting phosphorescence (red).

Because excitation leads to higher-energy vibronic and electronic states, but the return to the ground state occurs from S_1,V_0, fluorescence emission is always of longer wavelengths (red-shifted) than absorption/excitation. This difference between absorption and emission maxima is the *Stokes shift*. The solvent can also contribute to the Stokes shift (Figure 21.34). Polar solvent molecules such as water form a solvation shell around the solute to stabilize its charges. Absorption is a fast process during which bond lengths and molecule geometries do not change (Section 21.2.2, Figure 21.6). Solvating water molecules thus maintain their arrangement around the solute upon excitation, but the excited state is more polar than the ground state, and the water is not optimally arranged for stabilization of the excited state charges (Section 21.2.4). Immediately after excitation, the energy of the excited state is therefore higher in aqueous solutions compared to non-polar solvents. During the nanosecond lifetime of the excited state, the water molecules have ample time to re-orient and to optimally stabilize the excited state of the fluorophore. This process is called *solvent relaxation.* Fluorescence then occurs from the excited state that is optimally stabilized by the solvent. Again, the transition is very rapid, and the water molecules maintain their orientation during the electronic transition. As a consequence, the fluorophore returns to an electronic ground state of higher energy than in non-polar solvents. Solvent relaxation then leads to the original, solvated ground state. Polar solvents thus increase the energy difference between excitation and emission wavelengths and contribute to the Stokes shift.

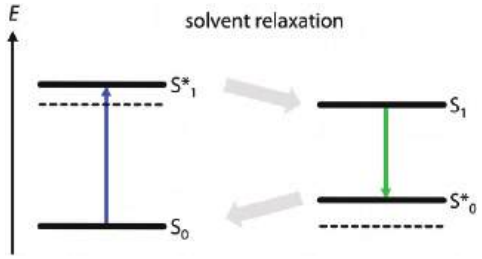


FIGURE 21.34: Solvent relaxation and Stokes shift. Polar solvents stabilize the charge distribution of solutes. Upon excitation, the charge distribution changes, but the arrangement of solvent molecules around the fluorophore is unaltered. The resulting excited state S^*_1 has a higher energy than in a non-polar solvent. Solvent relaxation (gray arrow) is the rearrangement of the solvent molecules to the energetically most favorable geometry for the electron distribution of the excited state (S_1). Relaxation of the fluorophore from S_1 to the ground state reverts the electron distribution. Again, the solvent molecules remain fixed, and the molecule reaches the S^*_0 state that is of higher energy than the original S_0 state. Solvent relaxation (gray) restores the original solvated ground state.

21.5.2 Instrumentation

In a standard fluorescence experiment using a conventional fluorescence spectrometer, the fluorophores in the sample are excited by continuous illumination, and the steady-state fluorescence emission intensity is measured. Two types of spectra can be measured, excitation and emission spectra. For *emission spectra*, the excitation wavelength is fixed, and the emission is recorded as a function of wavelength. *Excitation spectra* are measured the other way around: the excitation wavelength is varied and emission is recorded at a fixed wavelength, usually the emission maximum. A fluorescence excitation spectrum is equivalent to the absorption spectrum.

A standard fluorescence spectrometer (Figure 21.35) has a xenon lamp as a light source. An excitation monochromator selects the excitation wavelength. A second monochromator or filter selects the emission wavelength or wavelength range. Emitted light is detected by a photomultiplier. The emission/detection path is oriented at 90° with respect to the excitation path to avoid background from excitation light on the detector. Measured fluorescence spectra are always corrected for the spectrum of the solvent, recorded in a separate reference experiment. Fluctuations in the intensity of the lamp are detected by a reference diode and corrected by the instrument. The sensitivity of detectors is wavelength-dependent. Most commercial spectrometers automatically correct the spectra for this wavelength-dependent detection efficiency using an internal calibration file. For quantitative comparisons of spectra, it is important to be aware if these corrections are being made by the instrument used.

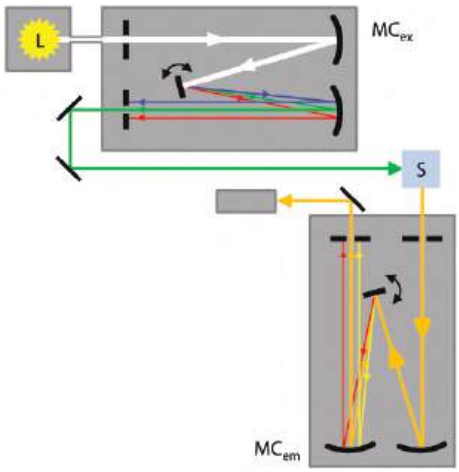


FIGURE 21.35: Fluorimeter. A light source (L) emits white light. The excitation monochromator (MC_{ex}) selects the desired excitation wavelength. Fluorescence light emitted from the sample (S) in a cuvette is detected at right angles. The emission light is split into different wavelengths by the emission monochromator (MC_{em}), and is detected by a photomultiplier.

21.5.3 Quantum yield and lifetime

The *quantum yield* ϕ of a fluorophore is the ratio of emitted photons relative to absorbed photons. ϕ depends on the kinetic competition (Section 8.2) of fluorescence with other events that depopulate the excited state without emission of light. The quantum yield can be calculated from the ratio of the rate constant for fluorescence, k_f, and the sum of all rate constants that lead away from the excited state, including those for non-radiative processes summarized here as k_{nr}:

$$\phi = \frac{k_f}{k_f + k_{nr}} \qquad \text{eq. 21.61}$$

In the absence of non-radiative processes, the quantum yield is $\phi = 1$, *i.e.* each excitation event leads to one fluorescence event. In practice, quantum yields will be lower due to the occurrence of non-radiative processes, such as internal conversion and inter-system crossing (Figure 21.33), quenching (Section 21.5.7), or energy transfer (Section 21.5.10). If these non-radiative processes are the dominant pathways for a system to relax to S_0, fluorescence may be totally suppressed ($\phi = 0$).

The *fluorescence lifetime* τ is the mean time the molecule spends in the S_1 state before returning to the ground state S_0 by emitting fluorescence. The lifetime is the inverse of the sum of all rate constants leading away from the S_1 state:

$$\tau = \frac{1}{k_f + k_{nr}} \qquad \text{eq. 21.62}$$

The *intrinsic fluorescence lifetime* τ_0 is defined as the lifetime in the absence of non-radiative processes, and is the inverse of the rate constant for fluorescence, k_f:

$$\tau_0 = \frac{1}{k_f} \qquad \text{eq. 21.63}$$

Combining eq. 21.62, eq. 21.63, and eq. 21.61, we obtain the relationship between quantum yield and lifetime as

$$\phi = \frac{\tau}{\tau_0} \qquad \text{eq. 21.64}$$

Quantum yields can be determined relative to reference substances (Box 21.8). Fluorescence lifetimes are measured in a lifetime spectrometer (see Section 21.5.9). We will see later that fluorescence lifetime and steady-state fluorescence emission are related, but the measurement of lifetimes provides additional information.

BOX 21.8: DETERMINATION OF QUANTUM YIELDS.

Quantum yields can be determined from fluorescence emission spectra in comparison with reference substances. To this end, the fluorescence emission spectrum of the reference and the sample are measured using the same instrument settings (excitation wavelength, detection and emission band width, photomultiplier voltage), and corrected for the instrument-specific wavelength dependence of sensitivity. From the areas under the emission band and the extinction coefficients of reference and sample at the excitation wavelength, the fluorescence quantum yield ϕ_s of the sample can be calculated (Figure 21.36). The quantum yield of fluorescein has been determined to high accuracy using thermal methods, and fluorescein in 0.1 M NaOH ($\phi = 0.95$) is often used as a reference. Magde *et al.* (2002) Photochem. Photobiol. 75: 327–334.

(*Continued*)

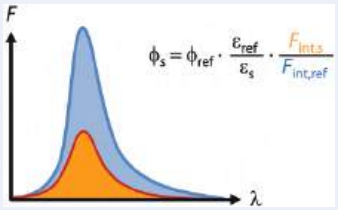


FIGURE 21.36: Determination of quantum yields. From the ratios of the integrated fluorescence F_{int} of a reference substance (ref, blue) and the sample (s, red), the quantum yield of the sample ϕ_s can be calculated according to the equation given. ε_{ref} is the extinction coefficient of the reference substance at the excitation wavelength, ε_s the extinction coefficient of the sample. ϕ_{ref} is the known quantum yield of the reference substance.

21.5.4 Fluorophores and fluorescent labeling

21.5.4.1 Biological fluorophores

Intrinsic fluorescence of biomolecules is restricted to aromatic amino acids, some cofactors, and unusual nucleobases. In proteins, the aromatic side chains of Trp, Tyr, and Phe absorb and emit in the UV spectral range (Figure 21.37). The quantum yields are ~0.2 (Trp), 0.14 (Tyr), and 0.04 (Phe). The extinction coefficients at the absorption maximum are ~5600 M^{-1} cm^{-1} for Trp (λ_{ex} = 280 nm), 1200 M^{-1} cm^{-1} for Tyr (λ_{ex} = 274 nm), and 200 M^{-1} cm^{-1} for Phe (λ_{ex} = 257 nm), and are a measure of the excitation efficiency (Table 21.2). The product of the extinction coefficient and the quantum yield is a measure of the relative fluorescence "strength" or sensitivity, and is highest for Trp (1120 M^{-1} cm^{-1}), followed by Tyr (168 M^{-1} cm^{-1}) and Phe (8 M^{-1} cm^{-1}). For fluorescence studies, Trp is therefore the most important and most useful amino acid. Trp shows the largest Stokes shift (70 nm, λ_{em} = 350 nm in aqueous environment). Trp also shows the largest dependence of its fluorescence on the environment, with large changes in quantum yield and strong shifts in emission wavelength to 320 nm and below in hydrophobic surroundings such as the interior of a protein. Therefore, Trp fluorescence is a powerful probe to follow protein folding/unfolding reactions (see Figure 21.55). In contrast, Tyr fluorescence shows a small Stokes shift of about 30–35 nm, which is almost independent of the polarity of the surroundings (λ_{em} = 305–310 nm). The quantum yield is also dependent on the environment, and Tyr fluorescence is therefore a suitable spectroscopic probe for protein folding and unfolding. Because of its low fluorescence compared to Trp and Tyr and its emission at 260 nm, Phe cannot be used as a fluorescence probe when Trp and/or Tyr are also present.

TABLE 21.2:
Absorption and fluorescence of aromatic amino acids.

amino acid	extinction coefficient ε (M^{-1} cm^{-1})	quantum yield ϕ	sensitivity $s = \varepsilon \cdot \phi$ (M^{-1} cm^{-1})
tryptophan	5600 (280 nm)	0.2	1120
tyrosine	1200 (274 nm)	0.14	168
phenylalanine	200 (254 nm)	0.04	8

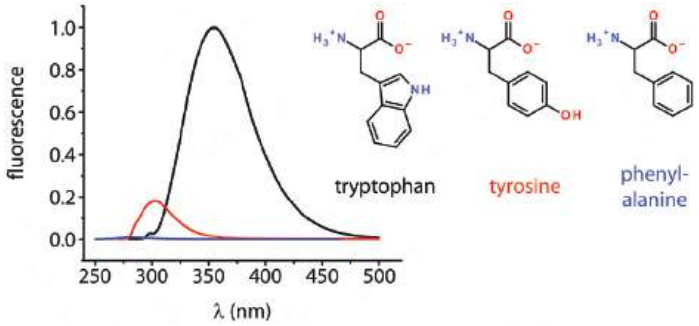


FIGURE 21.37: Fluorescence emission spectra of the aromatic amino acids tryptophan, tyrosine, and phenylalanine. Fluorescence emission spectra for tryptophan (λ_{ex} = 270 nm, ϕ = 0.12), tyrosine (λ_{ex} = 260 nm, ϕ = 0.13), and phenylalanine (λ_{ex} = 240 nm, ϕ = 0.022) in aqueous solution (arbitrary units). The structures of the amino acids are shown on the right. Data from Dixon *et al.* (2005) Photochem. Photobiol. 81(1), 212–213.

Several cofactors (see Figure 21.13) contain large aromatic systems that exhibit fluorescence. Examples are NADH (λ_{ex} = 340 nm, λ_{em} = 460 nm), with the oxidized form NAD^+ being non-fluorescent. In NADH, the fluorescence of the nicotinamide moiety is partially quenched due to stacking with the adenine ring (static quenching, see Section 21.5.7). NADH fluorescence often increases upon binding to proteins: binding of the dinucleotide in an extended conformation alleviates quenching, and energy can be transferred from Trp to the nicotinamide moiety.

Flavins and their derivatives FMN and FAD also show fluorescence (λ_{ex} ca. 450 nm, λ_{em} ca. 525 nm). As for NADH, the fluorescence is partially quenched in FMN and FAD. Notably, flavins show fluorescence in their oxidized form, and are non-fluorescent in the reduced state.



FIGURE 21.38: Pyridoxal phosphate (PLP). PLP is a common cofactor in transamination reactions. During the transamination reaction, an amino group from the first substrate is transferred to PLP and forms a Schiff base with the aldehyde group (orange). In the second step, this amino group is transferred to the second substrate of the reaction. The photophysical properties of PLP change during the catalytic cycle of transaminases.

The fluorescent pyridoxal phosphate (PLP; Figure 21.38) is an important cofactor in a multitude of reactions of amino acids (see Box 16.3). The absorption and emission maxima of PLP depend on its ligation state. During transamination reactions, the aldehyde group in PLP forms a Schiff base with the amino group of the substrate, which alters its photophysical properties.

Nucleotides and nucleic acids do not show fluorescence. Despite their strong absorbance, the canonical nucleobases have an extremely low quantum yield (on the order of 10^{-4}), resulting in fluorescence "strengths" of 0.04–0.3 M^{-1} cm^{-1}. Exceptions are rare: some modified bases such as 4-thiouracil and the Y base that occur in certain tRNAs show fluorescence in the UV range. Lipids and membranes lack conjugated electron systems and thus do not exhibit fluorescence.

21.5.4.2 Extrinsic fluorophores and their introduction into proteins, nucleic acids, and lipids

If the biomolecule of interest does not contain a fluorescent group, or if the fluorescent group is not in a suitable location to serve as a spectroscopic probe, fluorescent reporters can be introduced into proteins, nucleic acids, and lipids at the desired position by chemical modification. In addition to their limited occurrence, most intrinsic fluorophores exhibit fluorescence in the UV range, which is

experimentally less well accessible, and the extinction coefficients and quantum yields are typically moderate. For highly sensitive fluorescence measurements, fluorophores that absorb in the visible range of the spectrum and have high extinction coefficients and improved quantum yields are frequently employed. These range from rather small organic dyes to quantum dots and entire fluorescent proteins such as the green fluorescent protein (GFP; see Box 21.9). Fluorescent moieties that are introduced into the molecule of interest as reporter groups are also called *extrinsic fluorophores*.

Small organic fluorophores are often covalently attached to proteins using the thiol group of cysteines. A large number of organic dyes has been developed that cover the whole visible spectrum and show high extinction coefficients and quantum yields (Table 21.3) and absorption maxima that coincide with common laser lines. Many of these dyes are commercially available as maleimide or iodoacetamide derivatives that specifically react with thiols (Figure 21.39). Alternatively, succinimidyl ester or isothiocyanate derivatives are available to label amino groups.

TABLE 21.3:
Absorption and fluorescence properties of organic fluorophores.

fluorophore	λ_{max} (nm)	ε (M^{-1} cm^{-1})	λ_{em} (nm)	ϕ
fluorescein	495	75 000	521	0.90
AlexaFluor488	490	73 000	524	0.92
Atto488	500	90 000	520	0.80
Cy3	555	150 000	570	0.31
tetramethylrhodamine	546	90 000	580	0.70
AlexaFluor546	556	112 000	573	0.79
Cy5	646	250 000	662	0.20
Alexa647	650	270 000	665	0.33
Atto647	647	120 000	667	0.20
Cy7	750	199 000	773	0.30

Note that values are approximate only. Exact values depend on the solvent and the local environment of the fluorophores.

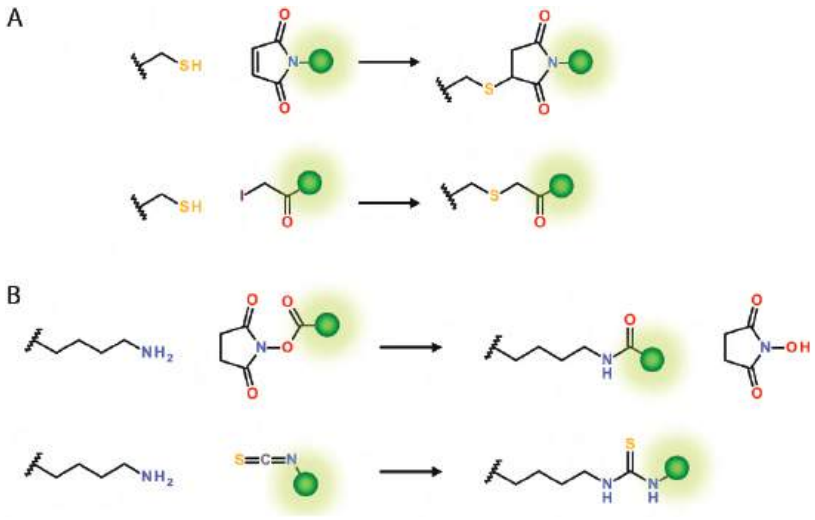


FIGURE 21.39: Fluorescent labeling of cysteines and lysines. A: Cysteines react with maleimides by addition to the double bond (top) and with iodoacetamides (bottom) by nucleophilic substitution. Maleimide or iodoacetamide derivatives of fluorescent dyes (green) are used for covalent attachment of fluorophores. B: Lysines can be covalently modified by reaction with succinimidyl esters (top) or isothiocyanates (bottom).

Because of the high natural abundance (~6%), labeling of lysines leads to the introduction of multiple fluorophores into the protein of interest. This approach is frequently used to generate fluorescently labeled antibodies. However, self-quenching of fluorescence (see Section 21.5.7) at high fluorophore densities influences their brightness. The optimal ratio of fluorophores per protein at which the fluorescence signal is maximal depends on the fluorophore.

The incorporation of non-natural amino acids by suppression of stop codons affords an increasing number of bio-orthogonal functional groups (see Section 18.1.5), such as keto or azido groups for site-specific modifications of proteins, using simple and efficient reactions (click chemistry, Figure 21.40). Functional groups for subsequent labeling can also be introduced enzymatically. One example is the formylglycine generating enzyme that can be used to genetically encode formyl groups (see Box 18.1). Farnesyl transferases have been used for the introduction of alkyne groups to a four amino acid sequence that can then be modified using click chemistry.

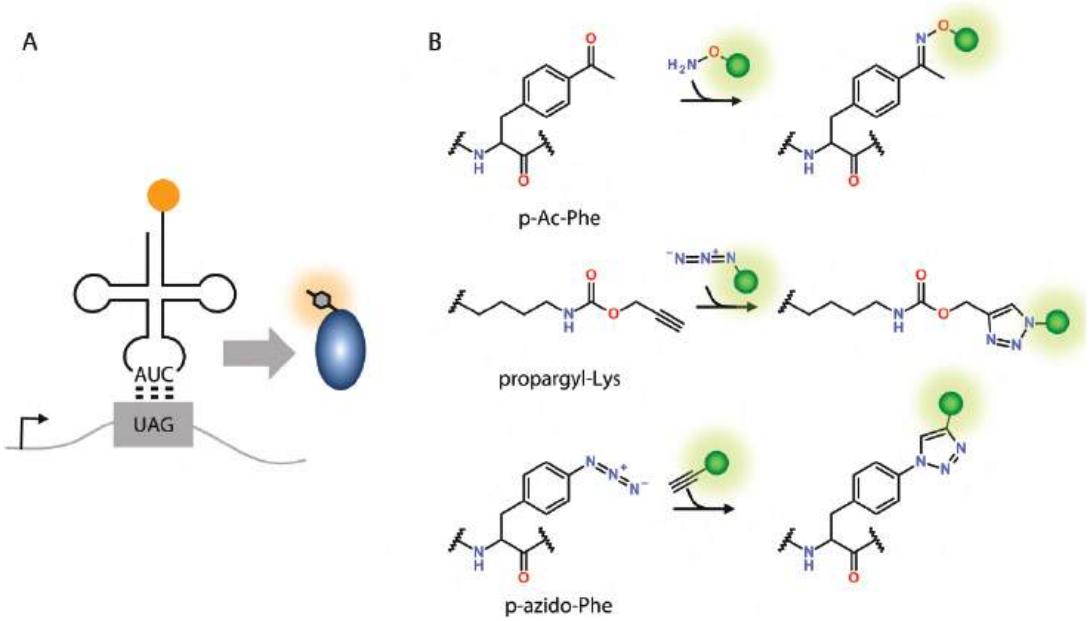


FIGURE 21.40: Click chemistry and labeling of non-natural amino acids. A: Internal stop codons are recognized by suppressor tRNAs. The suppressor tRNA can be charged with a non-natural amino acid (orange) that will be incorporated into the protein (blue, right). The non-natural amino acid contains a functional group that can be coupled to a fluorophore in a bio-orthogonal reaction. B: Examples of click chemistry reactions of non-natural amino acids: p-acetyl phenylalanine (p-Ac-Phe) reacts with hydroxylamine derivatives of dyes (green). Propargyl lysine and azido phenylalanine can undergo copper-mediated cycloadditions of azido- or alkyne-modified dyes, respectively.

A different approach for fluorescent labeling is based on the introduction of N- or C-terminal tags that are subsequently modified with a fluorophore *in vitro* or *in vivo* (Figure 21.41–Figure 21.43). Non-covalent fluorescent labeling is possible by complexation of a hexahistidine-tag with a fluorescent variant of Ni^{2+}-nitrilotriacetic acid (Figure 21.41). Similarly, tetracysteine motifs form non-covalent biarsenic complexes with fluorophores. Tags can also be used to introduce covalently attached fluorescent labels. The SNAP tag is a 20 kDa fragment of O^6-alkyl-guanine alkyltransferase, involved in DNA repair *in vivo*. The SNAP fragment uses benzylguanine as a substrate. Dyes (derivatized as N-hydroxy-succinimidylesters) can be coupled to its amino group to create a fluorescent benzylguanine derivative, which is then covalently coupled to a Cys of the SNAP tag (Figure 21.42). Similarly, the Halo tag® is derived from a modified haloalkane dehalogenase that has been engineered to accept synthetic chloroalkane ligands. The ligands are modified with fluorophores (or other reporters) and form a covalent bond with the Halo tag fused to the protein of interest. Alternatively, a transglutaminase recognition sequence (Q tag) can be introduced into the

protein of interest. Transglutaminase will then transfer a fluorescent probe with an amino group to a conserved Gln within this sequence. Finally, sortase is a bacterial cysteine peptidase that cleaves a C-terminal sorting signal from cell-wall-linked proteins. The five-amino-acid sorting signal (the sortase tag) can be coupled to fluorophores. Sortase recognizes the sortase tag and links it to the N-terminus of the target protein (Figure 21.43). Sortase can also be used for internal or C-terminal labeling. Many of these approaches are also employed for surface immobilization of biomolecules in surface plasmon resonance (Section 29.3.4) or fluorescence microscopy (Sections 25.4, 25.7).

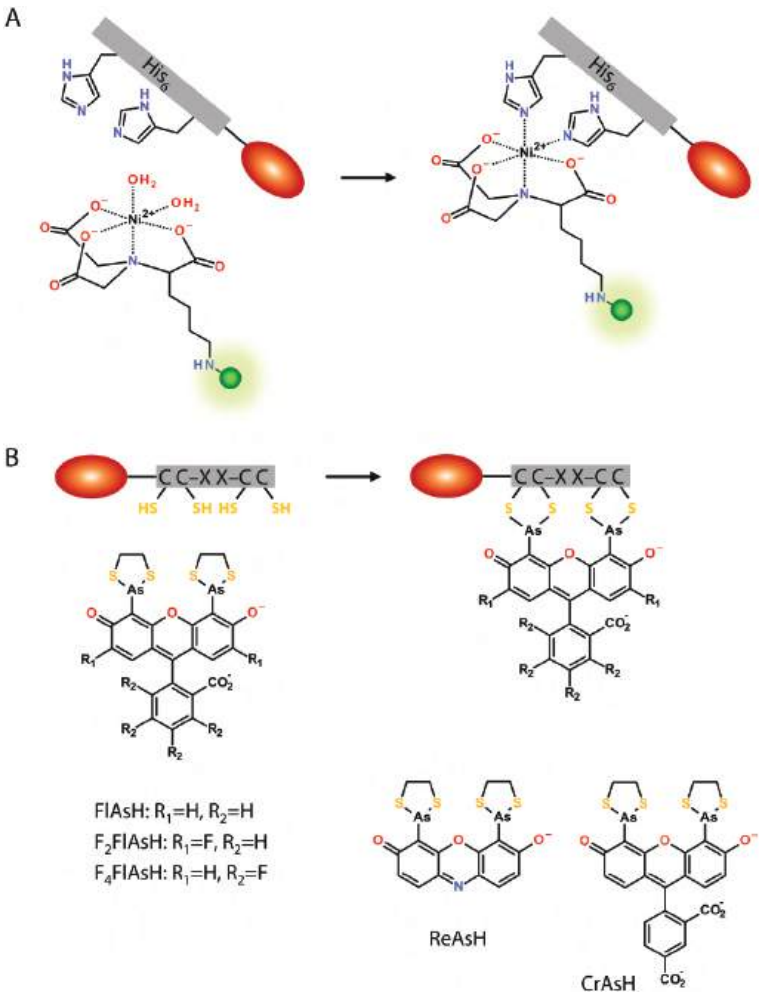


FIGURE 21.41: Introduction of tags and non-covalent fluorescent labeling. A: A His-tag (gray) fused to the protein of interest (orange) forms a non-covalent complex with Ni^{2+}-nitrilotriacetic acid (NTA). Using fluorescently modified NTA-derivatives, this interaction can be used to couple a fluorophore (green sphere) to a His-tagged protein. B: Motifs containing four cysteines (CCXXCC, gray) form non-covalent complexes with biarsenic fluorophore derivatives. The ethane dithiol forms of the biarsenic compounds are non-fluorescent, but complexation by the tetracysteine motif renders them fluorescent. Adding biarsenic complexes to the protein of interest (orange), fused to a tetracysteine motif, allows non-covalent fluorescent labeling of proteins. A number of different biarsenic compounds based on fluorescein or resorufin are available.

Expressed protein ligation is a procedure derived from the reaction sequence in protein splicing, in which a protein that is modified with a thioester at the C-terminus can react with a peptide (or protein) carrying an N-terminal cysteine (Figure 21.44). Thioester-modified proteins and proteins with N-terminal cysteines can be generated by production as fusion proteins with intein sequences (protein introns) and subsequent cleavage of the fusion protein. At the end of the ligation reaction, the two protein parts are linked *via* an authentic peptide bond. With a fluorophore-labeled peptide, this reaction can be used to introduce a fluorescent label into the protein of interest.

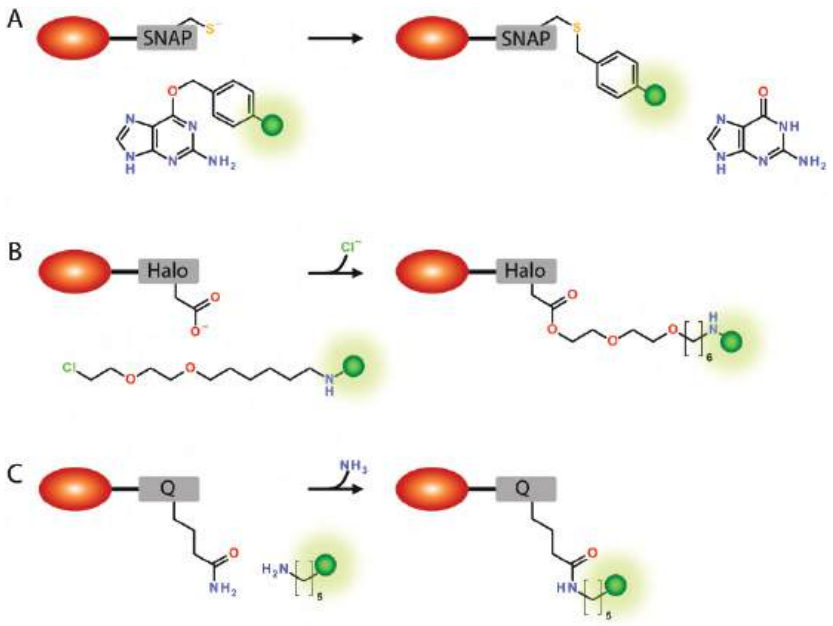


FIGURE 21.42: Introduction of tags for covalent fluorescent labeling. A: The SNAP tag (gray) is a 20 kDa fragment of O^6-alkyl-guanine alkyltransferase. The reaction of a cysteine within the tag with fluorescently labeled (green) benzylguanine can be exploited for covalent fluorescent labeling. B: The Halo tag® (gray) is derived from a haloalkane dehalogenase. An aspartate within the Halo tag forms a covalent adduct with fluorescently modified (green) chloroalkane ligands. C: The Q tag (gray) is a transglutaminase recognition sequence. Transglutaminase transfers fluorescently labeled (green) amines to a conserved glutamine within the tag. The target protein is represented in orange.

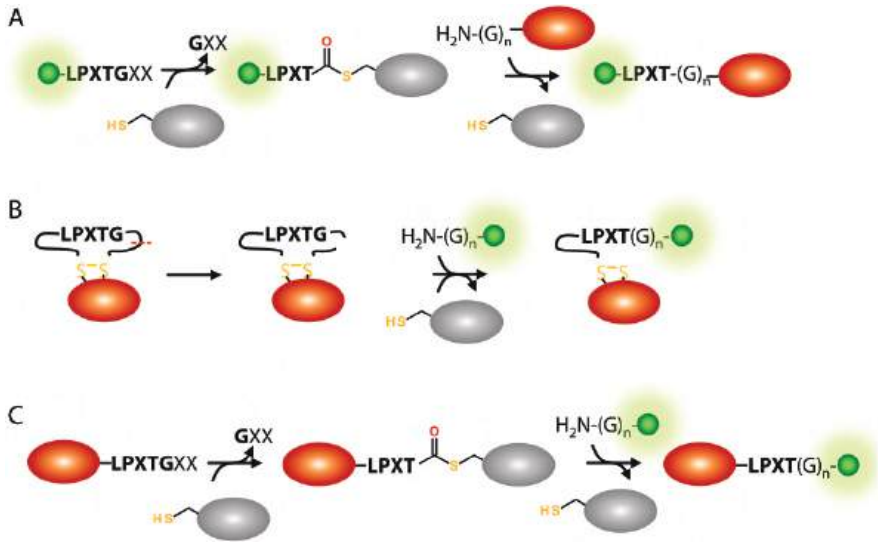


FIGURE 21.43: Sortase for N-terminal, internal, or C-terminal fluorescent labeling. A: Sortase (gray sphere), a bacterial cysteine peptidase that cleaves a C-terminal sorting signal from cell-wall-linked proteins, recognizes the five-amino-acid sortase tag (LPXTG, bold) and links it to the N-terminus of target proteins (orange). This reaction can be used for covalent fluorescent labeling (green sphere) of proteins at the N-terminus. B: For internal labeling, the sortase tag has to be introduced into the protein of interest (orange) within a loop that is closed by a disulfide bond. Cleavage by sortase generates a C-terminal sortase tag. Sortase then transfers this tag, including the protein of interest, to the N-terminus of a fluorescently modified (green) linker peptide. C: For C-terminal labeling, the sortase tag is fused to the C-terminus of the protein of interest (orange). Sortase cleaves the sortase tag between T and G, and couples the tagged protein of interest to the N-terminus of a fluorescently modified (green) linker peptide.

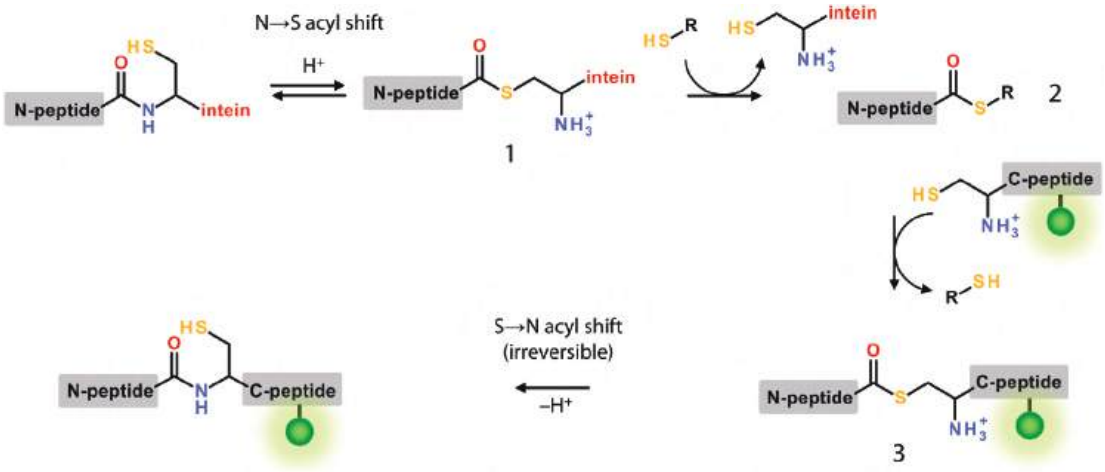


FIGURE 21.44: Expressed protein ligation. Two peptides (N-peptide, C-peptide) can be linked by a peptide bond *via* expressed protein ligation. The N-peptide is fused to an intein that has a cysteine at the N-terminus. The nucleophilic attack of the thiol group at the carbonyl carbon of the peptide bond causes an N→S acyl shift to a thioester (1). This thioester is further activated by addition of a thiol (R-SH) with a good leaving group R (2). This thioester reacts with the C-peptide that carries a cysteine at the N-terminus (3). An irreversible S→N acyl shift then leads to the final product, the two peptides linked by a peptide bond. Ligation of a recombinantly produced fusion protein of N-peptide with a fluorescently labeled synthetic C-peptide allows the introduction of fluorophores (green sphere) into larger proteins.

If their size is not an issue, proteins can be fluorescently labeled by fusion with fluorescent proteins. Starting from natural fluorescent proteins such as the green fluorescent protein (GFP) from the jellyfish *Aequorea victoria*, a full spectrum of fluorescent proteins has been developed with emission wavelengths that cover the whole visible spectrum (Box 21.9). While these fluorophores are genetically encoded and useful for fluorescence microscopy and *in vivo* localization of biomolecules, their large size of ~25 kDa and unstable fluorescence emission limit their applications.

BOX 21.9: FLUORESCENT PROTEINS.

Green fluorescent protein (GFP) is a 27 kDa protein from the jellyfish *Aequorea victoria* that exhibits green fluorescence after excitation with UV or blue light. GFP folds into a β-barrel structure (Figure 21.45). Air oxidation generates the fluorophore, formed by the side chains of a Ser-Tyr-Gly sequence. The fluorophore is located in the center of the protein, protected from solvent. Numerous GFP variants have been generated since its discovery, and it has been developed into a versatile genetic tool. The first variant was the S65T variant (eGFP) that showed enhanced fluorescence compared to wild-type GFP. Substitution of Y66 generated the cyan fluorescent protein (CFP), with a blueshift in absorbance and emission wavelength. More complex modifications of the fluorophore surroundings led to the yellow fluorescent protein (YFP). CFP and YFP are commonly used as FRET pairs (see Section 21.5.10) for *in vivo* studies. Naturally occurring red fluorescent proteins such as the dimeric DsRed isolated from the sea anemone *Discosoma striata* were the basis for a number of monomeric red fluorescent proteins with improved photophysical properties that cover the red spectral region.

(*Continued*)

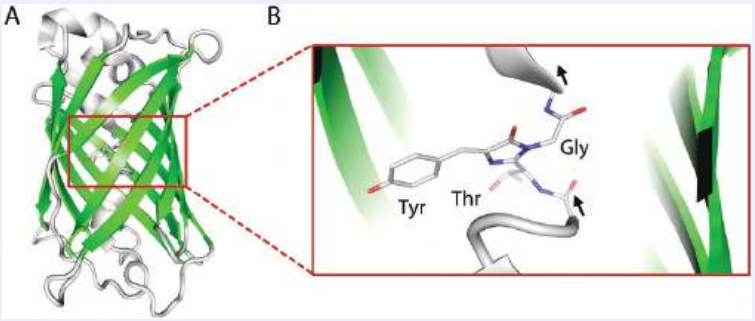


FIGURE 21.45: Structure of green fluorescent protein and its fluorophore. A: Green fluorescent protein (GFP) from the jellyfish *Aequorea victoria*. GFP forms a β-barrel structure, with the fluorophore buried inside the barrel. Upon excitation with blue light, GFP shows green fluorescence. B: The fluorophore in GFP is formed by oxidation and ring formation of the side chains of the three consecutive amino acids Ser65, Tyr66, and Gly67 (in wild-type GFP). The structure depicted here is the enhanced version with an S65T mutation.

Split versions of GFP can be used in complementation assays to investigate protein-protein or other biomolecular interactions. In these assays, a GFP fragment is fused to one potential interaction partner, the remainder of GFP is fused to the second partner. While the individual protein fragments do not exhibit fluorescence, complex formation *in vivo* allows the GFP fragments to recombine to a complete GFP that emits fluorescence. Tripartite GFP versions have also been used in *in vivo* interaction assays.

Organic fluorophores, although superior to intrinsic fluorophores, still pose problems for fluorescence measurements. A major drawback is their limited brightness and photostability and their tendency to undergo irreversible photobleaching (see Section 25.7.1, Box 25.7). *Quantum dots* are semiconductor nanocrystals whose diameter is similar in magnitude to their de Broglie wavelength, leading to a phenomenon called quantum confinement, and to special molecular properties. As a result of this quantum confinement, the bonding and anti-bonding orbitals of quantum dots are separated by a *band gap* that decreases with the size of the quantum dot (Figure 21.46). Absorption causes the transition of an electron from the bonding to the anti-bonding orbital, leaving a hole in the bonding orbital. Relaxation of the electron to the bonding orbital, and recombination with the hole, is then accompanied by fluorescence emission. As a consequence of the size-dependent band gap, the emission wavelength of quantum dots made from the same material depends on their size: larger quantum dots show red-shifted emission, the emission of small quantum dots is blue-shifted.

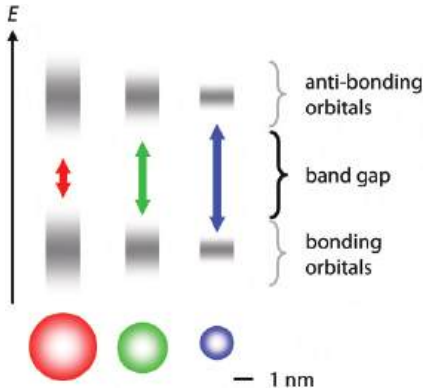


FIGURE 21.46: Quantum dots and the size dependence of the band gap. The bonding and anti-bonding orbitals of semiconductor nanocrystals (red, green, and blue spheres) are separated by a band gap. The larger the nanocrystal, the broader the energy band for each type of orbital, and the lower the energy required for excitation of electrons from a bonding to an anti-bonding orbital (double-headed arrows).

Nanocrystals can be coated with polymers, functionalized with streptavidin, and coupled to biotinylated molecules of interest. Compared to organic fluorophores, quantum dots show an increased brightness because of their higher extinction coefficients and comparable fluorescence quantum yields. They are photostable and show less photobleaching. However, quantum dots show frequent *blinking*, short excursions to a non-fluorescent state such as the triplet state. Blinking can be suppressed by adding Trolox, a vitamin E analog that acts as a triplet state quencher and rescues fluorophores from the non-fluorescent triplet state (see Box 25.7). The toxicity of semiconductor nanocrystals limits *in vivo* applications. Also, the name quantum dot is a bit deceiving when it comes to size. Typically, a functionalized quantum dot has a diameter of a few nanometers, corresponding to the size of a 30–50 kDa protein.

Fluorescent nucleotide analogs are powerful tools to study the interaction of proteins with nucleotides. In accordance with their widespread function in the cell, a wide range of nucleotides coupled to fluorescent reporter groups are commercially available. Examples are mant nucleotides that carry a N-methylanthraniloyl (mant) group that forms an ester bond with the 2'- or 3'-hydroxyl group of the ribose or deoxyribose (Figure 21.47).

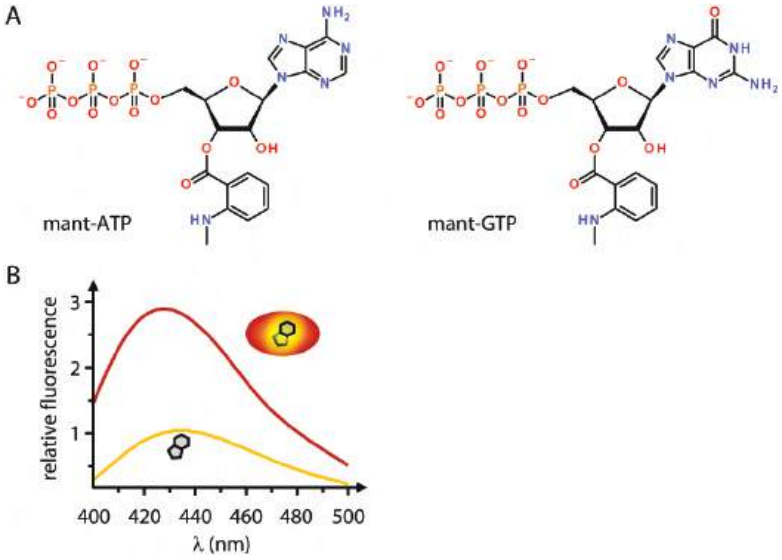


FIGURE 21.47: mant nucleotides as a probe for nucleotide binding. A: mant-ATP and mant-GTP are frequently used to study nucleotide binding to ATP- or GTP-binding proteins. B: Fluorescence of the mant group can be excited at 360 nm. The emission maximum is 440 nm. Upon binding to a protein, the mant fluorescence increases by 2–3-fold due to the alleviation of intermolecular quenching by the adenine/guanine base.

Similar to the static quenching (see Section 21.5.7) in NADH and FAD, the mant fluorescence of mant nucleotides in solution is quenched by the nucleobase that stacks onto the mant group. After specific binding of the nucleobase by a protein, stacking is no longer possible and the mant fluorescence increases, often by a factor of 2–3.

Fluorescent labeling of nucleic acids can be achieved by introducing organic fluorophores during solid-phase synthesis using modified building blocks. This approach is possible only for fluorophores that are stable under the conditions of the coupling and deprotection reactions. One example of such a probe is 2-aminopurine (2-AP; λ_{ex} = 310 nm, λ_{em} ca. 370 nm), an adenine analog that base-pairs with thymine, but also with cytosine. By incorporating 2-AP in place of adenine, a chemically very similar fluorescent reporter can be introduced into DNA. 2-AP shows strong fluorescence in solution, but its fluorescence is partially quenched due to base stacking when it is part of a DNA duplex. Hence, 2-AP serves as a spectroscopic probe for DNA structure and dynamics, and for the accessibility of nucleotides in nucleic acids. 2-AP fluorescence has been used to monitor base-flipping by DNA-modifying enzymes or unwinding of DNA or RNA duplexes by helicases (Figure 21.48).

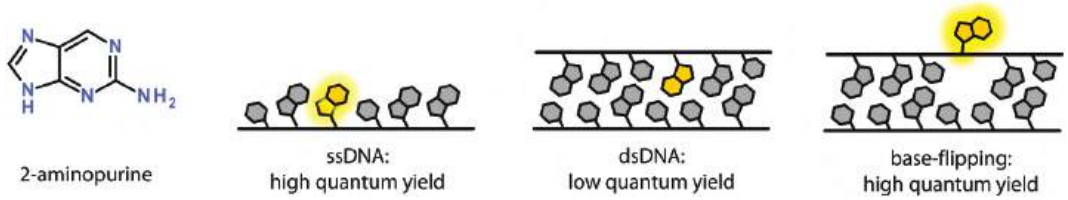


FIGURE 21.48: 2-aminopurine fluorescence as a probe for nucleic acid structure and dynamics. The fluorescence of 2-aminopurine is quenched by base stacking: its fluorescence is lower in double-stranded than in single-stranded DNA. The increase in fluorescence upon duplex separation can be exploited as a spectroscopic probe to monitor unwinding by DNA (or RNA) helicases. Some DNA modification enzymes modify bases while they are flipped out from the duplex. Base flipping alleviates stacking, and is associated with an increase in 2-AP fluorescence.

Fluorescent nucleotides have been generated by organic synthesis. Cytosine analogs with UV absorbance and fluorescence have been developed and introduced into DNA by chemical synthesis (Figure 21.49).

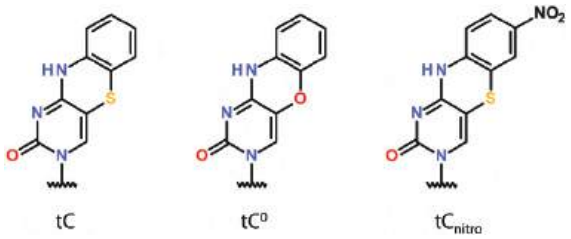


FIGURE 21.49: Fluorescent cytosine analogs. The cytosine analogs tC, tC^0, and tC$_{nitro}$ can be incorporated into DNA by solid-phase synthesis for fluorescence studies. tC and tC^0 exhibit fluorescence, whereas tC$_{nitro}$ is non-fluorescent, but can be used as an acceptor in FRET experiments (see Section 21.5.10).

As an alternative to introducing fluorophores into DNA directly during solid-phase synthesis, reactive groups can be introduced instead and modified with fluorophores after synthesis. This procedure allows the incorporation of fluorophores that would not withstand the harsh conditions during DNA synthesis. Longer DNA molecules (> 200 nucleotides) are not amenable to chemical synthesis. In this case, fluorophores can be introduced during PCR reactions with fluorescently modified primers (generated by solid-phase synthesis). Similarly, fluorescently labeled RNA can be generated enzymatically by *in vitro* transcription from a DNA template, using fluorescently modified starter oligonucleotides. Ligation techniques enable the covalent linkage of DNA and RNA fragments, chemically synthesized or enzymatically generated, to large nucleic acid molecules with fluorophores in the desired position(s). In structured RNAs, fluorescent labels have successfully been introduced by hybridizing fluorescently labeled oligonucleotides to unpaired bases in loop regions. A green fluorescent RNA molecule, called spinach, can be used as a genetically encoded green fluorescent tag for RNA labeling *in vivo*, in the same way as GFP for proteins.

Lipids can be modified with hydrophilic dyes at the polar head group, or with hydrophobic dyes introduced into the hydrophobic side chain without interfering with membrane formation, enabling fluorescent studies on lipid bilayers, micelles, and vesicles. A variety of such lipids is commercially available.

21.5.5 Applications

21.5.5.1 Fluorescence as a probe for binding: equilibrium titrations

Fluorescence is frequently used as a probe to monitor binding. In some cases, changes in Trp fluorescence can be employed to monitor ligand binding to proteins, or protein-protein interactions. More often, ligands labeled with extrinsic fluorophores are used as reporters for binding. K_d values for the

dissociation of macromolecule-ligand complexes can be determined in *equilibrium titrations* of the fluorescently labeled partner (typically the ligand) with the unlabeled macromolecule (Figure 21.50).

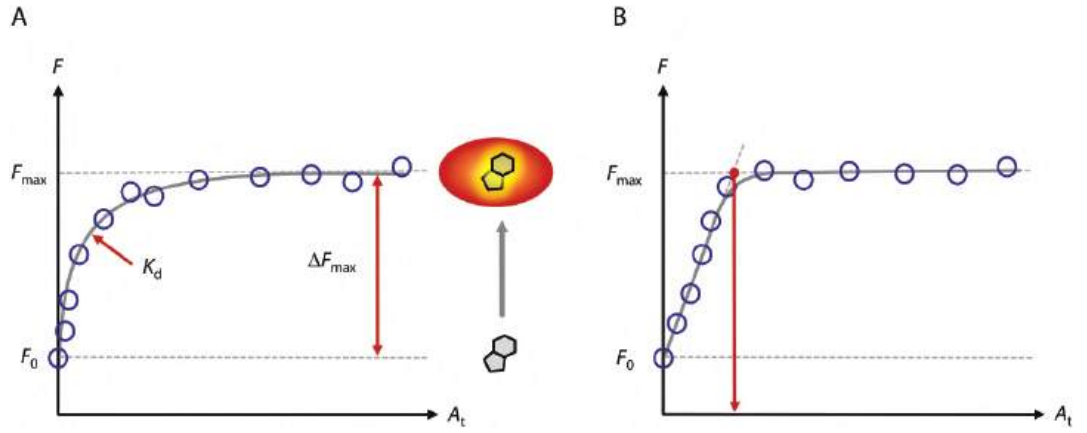


FIGURE 21.50: Fluorescence equilibrium titration. A: Fluorescence equilibrium titration. A fluorescent or fluorescently labeled ligand, provided at a concentration in the range of the expected K_d value, is titrated with its binding partner A. The fluorescence increase upon complex formation serves as a spectroscopic probe for binding. The binding curve obtained (blue circles) can be described by eq. 21.69 (gray curve) to determine the dissociation constant K_d of the complex. B: Active-site titration. The fluorescently labeled ligand is provided at a concentration more than 10-fold above the K_d value of the complex, such that nearly all of the macromolecule that is added in each titration step is bound. From the equivalence point (red arrow) of the binding curve, the stoichiometry of binding can be determined.

For a simple binding equilibrium of a fluorescent ligand B that binds to a biological macromolecule A to form an AB complex

$$A + B \underset{K_d}{\rightleftharpoons} AB \qquad \text{scheme 21.1}$$

we can formulate the equilibrium constant and the mass conservation, and obtain the concentration of the AB complex as a function of the total concentrations A_t and B_t as

$$[AB] = \frac{A_t + B_t + K_d}{2} - \sqrt{\left(\frac{A_t + B_t + K_d}{2}\right)^2 - A_t B_t} \qquad \text{eq. 21.65}$$

(see Section 2.2). We now have to relate the concentration of the AB complex to our spectroscopic signal. Let us assume that we measure the fluorescence of B, which changes upon AB complex formation. To determine the dissociation constant of the AB complex, we use a fixed concentration B_0 in the range of the K_d value we expect. In a *titration* experiment, we increase the concentration of A in a stepwise manner, and obtain a binding curve from the fluorescence of B as a function of A_t (Figure 21.50). The change in fluorescence of B upon formation of the AB complex is proportional to the degree of binding ν:

$$\nu = \frac{[AB]}{B_t} \qquad \text{eq. 21.66}$$

We express the total change in fluorescence ΔF for each step of the titration as the product of the degree of binding and the maximum fluorescence change ΔF_{max} upon saturation (ν = 1) at the end of the titration:

$$\Delta F = \Delta F_{max} \frac{[AB]}{B_t} \qquad \text{eq. 21.67}$$

Our fluorescence signal F at any given concentration of A then becomes

$$F = F_0 + \Delta F_{max} \frac{[AB]}{B_t} \qquad \text{eq. 21.68}$$

F_0 is the fluorescence of free B in the absence of A, and thus the starting value of the titration. In combination with eq. 21.65 we obtain

$$F = F_0 + \Delta F_{max} \cdot \frac{\frac{A_t + B_t + K_d}{2} - \sqrt{\left(\frac{A_t + B_t + K_d}{2}\right)^2 - A_t B_t}}{B_t} \qquad \text{eq. 21.69}$$

Eq. 21.69 thus describes the fluorescence F as a function of the total concentrations B_t and A_t and the dissociation constant K_d. Since we know A_t and B_t for each point of our titration, we can use eq. 21.69 to analyze the titration data and to determine K_d by non-linear least squares fitting. It is important to note that equilibrium titrations can be performed using any spectroscopic signal that changes linearly with complex formation, and K_d values can be determined from these titration curves by the same principle as outlined here.

If the ligand is provided at concentrations above the K_d, the binding equilibrium will be strongly on the side of the complex in each step of the titration because of the high ligand concentration. Virtually all of the added macromolecule will bind to a ligand molecule, giving rise to a nearly linear change in fluorescence (Figure 21.50). Saturation is reached when all ligand is bound, and further addition of macromolecule does not change the fluorescence. From the equivalence point, the stoichiometry of binding can be determined. Such an *active-site titration* is a convenient and often much more sensitive alternative to the determination of binding stoichiometries by isothermal titration calorimetry (Section 30.1).

In a fluorescence titration, we determine the K_d value for a complex of the fluorescently labeled ligand and its binding partner. What we really want to know is the K_d value of the complex with a non-fluorescent, natural ligand, however. Again, fluorescence equilibrium titrations are a useful approach. Once we have determined the K_d value of the complex with the fluorescent ligand, for example a fluorescently labeled nucleotide, we can perform a *competitive* or *displacement titration*, in which the fluorescent ligand and the non-fluorescent, authentic ligand are present at the same time and compete for the same binding site. We can formally distinguish between competitive and displacement titrations (Figure 21.51). In competitive titrations, the fluorescent ligand is titrated with the molecule of interest in the absence and presence of different concentrations of the non-fluorescent ligand. From the effect of the non-fluorescent ligand on the apparent K_d value, the true K_d value for the complex of the molecule with the non-fluorescent ligand is determined. Displacement titrations start from the complex of the molecule of interest with the fluorescent ligand. The fluorescent ligand is then displaced by adding increasing concentrations of the non-fluorescent ligand. As the fluorescent ligand becomes liberated, its fluorescence will return to the characteristic value of the unbound form. The result of these two approaches should be identical, provided that the reaction is in equilibrium for each titration point.

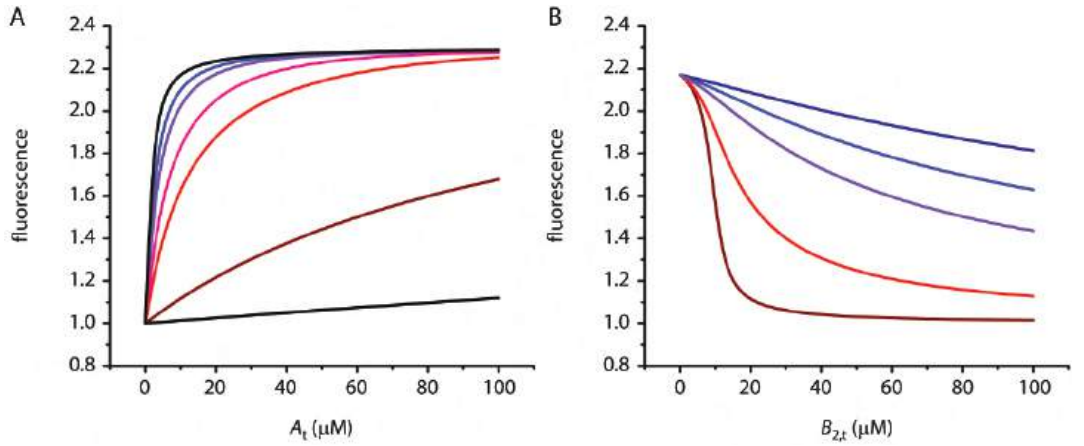


FIGURE 21.51: Fluorescence equilibrium competitive and displacement titration. A: Competitive titrations of a fluorescent ligand B_1 with a macromolecule A in the absence of competitor (black), and in the presence of increasing concentrations of a non-fluorescent competitor B_2. $B_{1,t}$ = 1 μM, K_{d1} = 1 μM, black (top): no competitor, blue: 10 μM, purple: 20 μM, pink: 50 μM, red: 100 μM, brown: 1 mM, black: 10 mM competitor. B: Displacement titrations of complexes of the macromolecule A and the fluorescent ligand B_1 with increasing concentrations of the non-fluorescent ligand B_2. K_{d1} = 1 μM, A_t = 10 μM, $B_{1,t}$ = 1 μM, brown: K_{d2} = 0.1 μM, red: K_{d2} = 1 μM, purple: K_{d2} = 5 μM, blue: K_{d2} = 10 μM, dark blue: K_{d2} = 20 μM.

In competitive and displacement titrations two equilibria have to be considered at each point of the titration. The first is binding of the fluorescently labeled ligand B_1 with its dissociation constant K_{d1}:

$$A + B_1 \underset{K_{d1}}{\rightleftharpoons} AB_1 \qquad \text{scheme 21.2}$$

The second equilibrium is binding of the non-fluorescent ligand B_2 with the dissociation constant K_{d2}:

$$A + B_2 \underset{K_{d2}}{\rightleftharpoons} AB_2 \qquad \text{scheme 21.3}$$

The equilibrium constants K_{d1} and K_{d2} are defined as

$$K_{d1} = \frac{[A][B_1]}{[AB_1]} \qquad \text{eq. 21.70}$$

and

$$K_{d2} = \frac{[A][B_2]}{[AB_2]} \qquad \text{eq. 21.71}$$

Both eq. 21.70 and eq. 21.71 have to be fulfilled at each point of the titration. To derive an expression for the fluorescence as a function of the concentration of the non-fluorescent ligand B_2, we again need to combine the laws of mass action with the mass conservation for B_1 and B_2. From these equations, we obtain an expression for the AB_1 complex as a function of the total concentrations A_t, $B_{1,t}$, and $B_{2,t}$. This time, the result is a cubic equation for the concentration of the AB_1 complex that provides our spectroscopic signal (Box 21.10). The solutions of the cubic equation are used to describe the fluorescence signal as a function of A_t, $B_{1,t}$ and $B_{2,t}$, and K_{d1} and K_{d2} in analogy to eq. 21.68 and eq. 21.69. Using this expression, K_{d2} is determined from displacement curves by non-linear least squares fitting. K_{d1} is determined independently in a separate titration experiment and is kept constant during the fitting procedure.

BOX 21.10: DETERMINING K_d VALUES FROM DISPLACEMENT TITRATIONS.

To derive an expression for the fluorescence signal as a function of the concentrations of A, B_1 and B_2, and the two K_d values, we again have to combine the mass conservation for B_1 and B_2 and the equilibrium constants K_{d1} and K_{d2}. A cubic equation for the concentration of the AB_1 complex is obtained:

$$0 = [AB_1]^3 + a \cdot [AB_1]^2 + b \cdot [AB_1] + c \qquad \text{eq. 21.72}$$

with the coefficients

$$a = \frac{A_t(K_{d2} - K_{d1}) + B_{1,t}(2K_{d2} - K_{d1}) + B_{2,t}K_{d1} - K_{d1}^2 + K_{d1}K_{d2}}{K_{d1} - K_{d2}} \qquad \text{eq. 21.73}$$

$$b = \frac{A_t B_{1,t}(K_{d1} - 2K_{d2}) - B_{1,t}^2 K_{d2} - B_{1,t}K_{d1}(B_{2,t} + K_{d2})}{K_{d1} - K_{d2}} \qquad \text{eq. 21.74}$$

and

$$c = \frac{A_t B_{1,t}^2 K_{d2}}{K_{d1} - K_{d2}} \qquad \text{eq. 21.75}$$

This cubic equation has three solutions. For

$$Q \equiv \frac{a^2 - 3b}{9} \qquad \text{eq. 21.76}$$

and

$$R \equiv \frac{2a^3 - 9ab + 27c}{54} \qquad \text{eq. 21.77}$$

and the condition

$$Q^3 - R^2 \geq 0 \qquad \text{eq. 21.78}$$

the three solutions are

$$[AB_1]_1 = -2\sqrt{Q}\cos\left(\frac{\theta}{3}\right) - \frac{a}{3} \qquad \text{eq. 21.79}$$

$$[AB_1]_2 = -2\sqrt{Q}\cos\left(\frac{\theta + 2\pi}{3}\right) - \frac{a}{3} \qquad \text{eq. 21.80}$$

and

$$[AB_1]_3 = -2\sqrt{Q}\cos\left(\frac{\theta + 4\pi}{3}\right) - \frac{a}{3} \qquad \text{eq. 21.81}$$

with

$$\theta = \arccos\left(\frac{R}{\sqrt{Q^3}}\right) \qquad \text{eq. 21.82}$$

(*Continued*)

For

$$Q^3 - R^2 < 0 \qquad \text{eq. 21.83}$$

eq. 21.72 has only a single solution:

$$[AB_1] = -\mathrm{sgn}(R)\left(\left(\sqrt{R^2 - Q^3} + |R|\right)^{\frac{1}{3}} + \frac{Q}{\left(\sqrt{R^2 - Q^3} + |R|\right)^{\frac{1}{3}}}\right) - \frac{a}{3} \qquad \text{eq. 21.84}$$

The solutions for $[AB_1]$ are used to express the fluorescence signal as a function of A_t, $B_{1,t}$, $B_{2,t}$, K_{d1}, and K_{d2} based on eq. 21.68:

$$F = F_0 + \Delta F_{max} \frac{[AB_1]}{B_{1,t}} \qquad \text{eq. 21.85}$$

The resulting equations can be implemented in data analysis software to determine K_{d2} from displacement curves by non-linear least squares fitting. K_{d1} is kept constant during the fitting procedure.

Fluorescence is also used as a probe in the determination of dissociation constants by microscale thermophoresis, a technique that is based on the different response of free and bound forms of molecules to a temperature gradient (Box 21.11).

BOX 21.11: DETERMINING K_d VALUES BY MICROSCALE THERMOPHORESIS.

Microscale thermophoresis (MST) is a method to analyze molecular interactions based on different responses of free and bound states of a molecule to a temperature gradient. In general, lighter particles (free form) move toward regions with higher temperatures, whereas heavier particles (bound form) move toward regions with lower temperature. While the process is well understood in the gas phase, its molecular origin in liquid mixtures is less clear. Here, size, charge, and the hydration shell of the molecules determine the magnitude and direction of the *thermophoretic effect*.

MST experiments are performed on solutions within glass capillaries (Figure 21.52). The molecule of interest is detected through its intrinsic fluorescence ("label-free") or through an extrinsic fluorophore attached to one of the interacting partners, enabling the determination of K_d values ranging from pM to mM. There are no restrictions with respect to the molecular mass of the interacting partners. The temperature gradient is generated by illumination of a small area (tens of µm) with an infrared laser; typical differences in temperature are between 2 K and 6 K. The thermophoretic effect then leads to a depletion or accumulation of the molecules in this area (Figure 21.52). The ratio of the concentrations in the hot and cold regions of the capillary, c_{hot} and c_{cold}, depends on the Soret coefficient S_T:

$$\frac{c_{hot}}{c_{cold}} = e^{-S_T \Delta T} \qquad \text{eq. 21.86}$$

(*Continued*)

where ΔT is the magnitude of the temperature gradient.

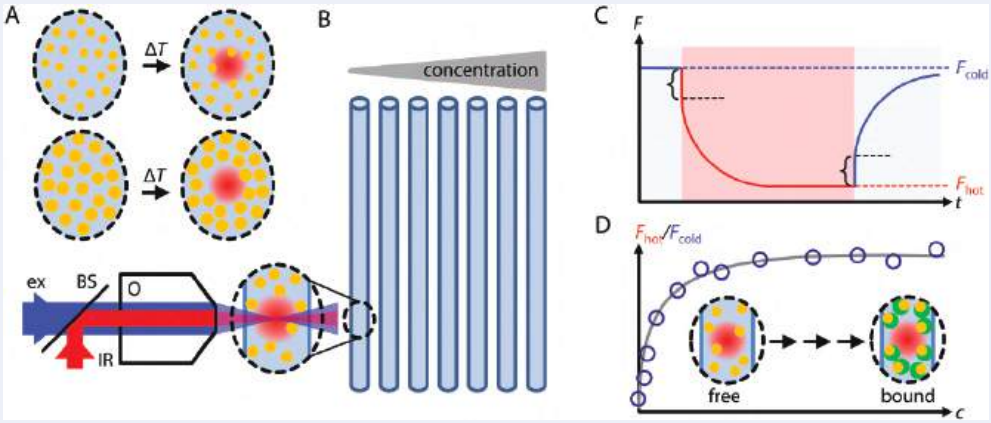


FIGURE 21.52: Microscale thermophoresis. A: Principle of thermophoresis. In the absence of a temperature gradient (left), molecules are equally distributed over the volume. When a temperature gradient is generated (right, red area), molecules tend to move toward (positive thermophoresis) or away from the hotter area (negative thermophoresis, shown here). The effect is larger for larger molecules (bottom). B: Measurement principle. A solution of the fluorescent interaction partner at a fixed concentration and increasing concentrations of the non-fluorescent binding partner is filled into capillaries. The fluorescence of the solutions is measured by scanning microscopy before (top) and after (bottom) the temperature gradient has been applied. ex: excitation light (blue) for fluorescence detection, IR: infrared laser (red) for generation of the temperature gradient ΔT, BS: beam splitter, O: objective. C: Fluorescence signal. F_{cold}: fluorescence intensity before the temperature gradient, F_{hot}: fluorescence intensity after the temperature gradient. The short dashed lines indicate the rapid signal change caused by the temperature dependence of fluorescence. D: The ratio F_{hot}/F_{cold} as a function of the concentration of the interaction partner gives the binding isotherm, which can be analyzed to determine the K_d value. The insets show the difference in thermophoresis for the free (orange) and bound (orange/green) form of the detected binding partner.

The fluorescence intensity is used as a read-out for the concentration of the fluorescent species in the hot and cold regions. First, the starting fluorescence (F_{cold}) before the application of the temperature gradient is measured. Then, the temperature gradient is generated. The temperature dependence of fluorescence leads to a rapid decrease of the intensity in the hot area, followed by a slow decrease (or increase) caused by thermophoretic depletion (or accumulation) of molecules in the heated area. The new equilibrium fluorescence value is recorded as F_{hot} (Figure 21.52). K_d values are determined by performing such measurements in parallel on a set of capillaries containing a fixed concentration of the interaction partner used for detection, and increasing concentrations of the other partner. The fluorescence traces from individual capillaries are normalized to the initial fluorescence value F_{cold}. The binding isotherm is obtained from the changes in fluorescence ratio F_{hot}/F_{cold} as a function of the concentration of the unlabeled partner (Figure 21.52), which can then be analyzed for the K_d value in the same way as a fluorescence equilibrium titration (eq. 21.69).

Fluorescence can also be measured as a function of time, and used as a spectroscopic probe to measure association and dissociation rates (see Chapter 28). From these experiments, macroscopic and microscopic rate constants can be extracted (Chapters 7–9, 11, 12).

21.5.5.2 Fluorescence as a probe for the chemical micro- and macro-environment

Fluorescence reports on the micro-environment of the fluorophore within a biomolecule, and can therefore be used as a probe for structural changes. Due to its high sensitivity for the local chemical

environment, Trp fluorescence is a valuable probe to monitor folding and unfolding of proteins (Figure 21.53). 1-anilinonaphtalene-8-sulfonate (ANS) binds to exposed hydrophobic regions of proteins, which is accompanied by an increase in its fluorescence. ANS has been used to monitor their folding and unfolding and to detect partially folded intermediates. The nucleobase 2-aminopurine (2-AP) is a probe for the local environment in the DNA, with high fluorescence when it is part of a single-strand, and low fluorescence within a double-stranded DNA (Figure 21.48). 2-AP fluorescence reports on structural changes of nucleic acids, and has been used to study folding of RNA molecules and to monitor structural changes in RNA and DNA. A similar principle is used in the thermal shift assay (Box 18.8).

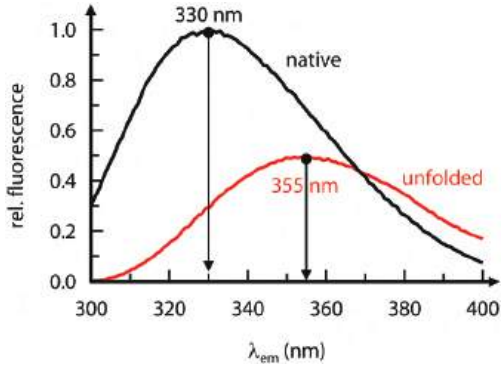


FIGURE 21.53: Tryptophan fluorescence as a probe to monitor protein unfolding. Fluorescence emission spectra of a native protein (black) and the same protein in the unfolded state (red). The fluorescence signal decreases due to collisional quenching by water, and the maximum shifts from 330 nm to 355 nm due to the solvent exposure of tryptophan side chains upon unfolding and the resulting more polar environment (see Figure 21.9).

Fluorescence can also be used as a probe for the macroscopic environment. Fluorescence sensors bind to an analyte and respond to its presence by a change in fluorescence. Thereby, they can report on pH values or concentrations of ions and metabolites (Box 21.12).

BOX 21.12: FLUORESCENCE SENSORS.

Fluorescein shows a fluorescence emission that is dependent on the pH, and can therefore function as a pH sensor. Fluorescein exists in two fluorescent protonation states, the mono-anionic and the di-anionic forms that equilibrate with a pK_A value of ~6.5. The two forms show different absorption and emission wavelengths, and fluorescein can therefore be used as a ratiometric pH probe by measuring emission after excitation at 450 nm (maximum of the mono-anion) and 495 nm (maximum of the di-anionic form). Fluorescence intensities depend on the concentration of the fluorophore, which is difficult to control in *in vivo* applications. Ratiometric approaches, in which the concentration-independent ratio of the fluorescence emission at two wavelengths is measured, are therefore often preferred. Fluorescence sensors for Ca^{2+} and Mg^{2+} ions are chelating molecules that form a high-affinity complex with the respective ion, accompanied by a change in fluorescence. One example is the dye Fura-2, which exhibits a shift in excitation wavelength in the presence of Ca^{2+} and is used as a ratiometric Ca^{2+} sensor. Other fluorescence properties such as fluorescence lifetimes (Section 21.5.9), anisotropy (Section 21.5.8), quenching (Section 21.5.7), or FRET (Section 21.5.10) can also be used for sensing.

21.5.5.3 Fluorescence and imaging: fluorescence recovery after photobleaching

Fluorescence is commonly used in imaging applications. Fluorescence microscopy, which will be discussed in detail in Chapter 25, allows fluorescent or fluorescently labeled biomolecules to be localized and tracked over time (see Sections 25.2, 25.7). Fluorescence microscopy can be employed to monitor diffusion of fluorescently labeled proteins or lipids in a membrane. In *fluorescence recovery after photobleaching* (FRAP) experiments, a membrane area (about 1 or 2 µm in diameter) containing fluorescently labeled lipids or proteins is illuminated with laser light in a microscope, leading to fluorophore excitation and steady-state emission at a constant intensity. At time t_0, the light intensity is increased by several orders of magnitude, leading to irreversible photodestruction (bleaching) of the fluorophore and a drop in emitted fluorescence. After the bleaching pulse, the excitation intensity is reduced again to its previous level. The fluorescence signal in the bleached area will now increase over time until it reaches nearly the original value. Fluorescence recovery is caused by fluorescent molecules that re-enter the observed membrane area by lateral diffusion (Figure 21.54). The increase in fluorescence $F(t)$ follows a single exponential time-course:

$$F(t) = F(t_0) + \left(F(t_\infty) - F(t_0)\right)\exp\left(-k_{\text{diff}}t\right) \qquad \text{eq. 21.87}$$

where $F(t_0)$ is the initial fluorescence after the bleaching pulse, $F(t_\infty)$ the final recovered fluorescence value that equals the fluorescence before the bleaching pulse, and k_{diff} the rate constant of the exponential fluorescence recovery. From k_{diff}, the lateral diffusion constant D_{lateral} of the molecule of interest can be determined:

$$D_{\text{lateral}} = \frac{r^2}{4} \cdot k_{\text{diff}} \qquad \text{eq. 21.88}$$

r is the radius of the bleached area. With this method, diffusion coefficients between 10^{-7} and 10^{-11} cm^2 s^{-1} (Section 4.1) can be measured.

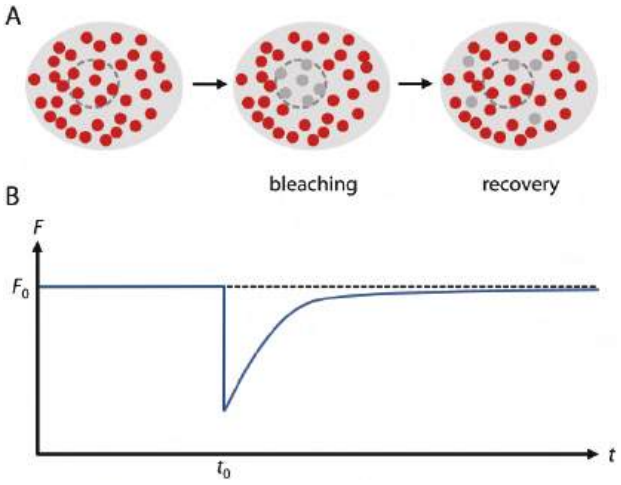


FIGURE 21.54: Fluorescence recovery after photobleaching (FRAP). A: Fluorescent proteins (or lipids, red) are evenly distributed in a membrane. At t_0, the area outlined in gray is irradiated with laser light to bleach the fluorophores in this region (gray). Over time, the fluorescently labeled molecules will redistribute evenly due to diffusion. B: The diffusion constant of the molecules can be determined from the time dependence of fluorescence recovery.

21.5.6 Potential pitfalls

Fluorescence experiments require a careful selection of measurement parameters such as excitation and emission wavelength, bandwidth of excitation (and emission), and fluorophore concentration. The emission is recorded at the wavelength with the maximum change in intensity during the

process under investigation. This wavelength can be determined from difference spectra that are calculated by subtracting the emission spectrum of the initial state from the spectrum of the final state (Figure 21.55). Maxima in the difference spectrum correspond to wavelengths with maximum increase in fluorescence, minima occur at the wavelength with maximum signal decrease.

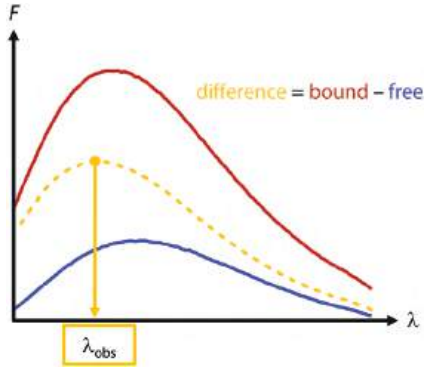


FIGURE 21.55: Difference spectra and signal maximization. Before performing an equilibrium fluorescence titration, the signal should be optimized. The maxima and minima of a difference spectrum (orange) between the final state (bound ligand, red) and the initial state (free ligand, blue) show at which wavelength the maximal increase/decrease in fluorescence is observed.

Photobleaching of fluorophores, leading to a decrease in the fluorescence signal with time, is a common problem in fluorescence experiments. An observed fluorescence decrease because of photobleaching can easily be misinterpreted as binding or as a structural change. Wherever possible, fluorescence experiments should preferentially be designed such that an increase in fluorescence can be monitored. If this is not possible, photobleaching can be minimized by using the lowest possible setting for the width of the excitation slit and thus for the excitation bandwidth, and by closing the shutter between individual measurements. Control experiments without addition of binding partners help to distinguish photobleaching from signal changes due to binding.

Fluorescence measurements have to be performed at sufficiently low concentrations to avoid inner-filter effects (Box 21.13, Figure 21.56). Otherwise, fluorescent molecules farther down the excitation pathway within the sample will experience a lower light intensity and show less fluorescence than molecules that are at the beginning of the light path and experience the full excitation intensity I_0. If an inner-filter effect occurs, the fluorescence intensity does not increase linearly with the concentration of the fluorophore, which is a key requirement for many biophysical applications of fluorescence spectroscopy. Whenever possible, the inner-filter effect should be avoided by limiting concentrations such that the absorption of the sample does not exceed 0.05. In other cases, an approximate correction of the measured values for the inner-filter effect is possible (Box 21.13, eq. 21.91).

BOX 21.13: INNER-FILTER EFFECT AND CORRECTION OF MEASURED FLUORESCENCE INTENSITIES.

At high chromophore concentrations a significant fraction of the excitation light is absorbed by chromophores at the beginning of the optical path. Molecules farther down toward the center of the cuvette will experience a lower light intensity, and hence produce a decreased emission intensity. This phenomenon is termed the *primary inner-filter effect*. Absorption of the emitted light while traveling from its origin in the center of the cuvette through the sample will cause a further decrease in the measured fluorescence intensity (*secondary inner-filter effect*). Both processes are referred to collectively as the *inner-filter effect*

(*Continued*)

(Figure 21.56). If the inner-filter effect cannot be avoided, measured fluorescence intensities can be corrected for the attenuation: According to the Lambert-Beer law, the absorption until the light has reached the center of the cuvette is half of the absorption over the entire path length ($d_{center} = d/2$).

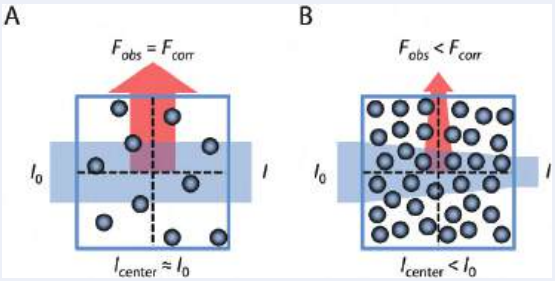


FIGURE 21.56: Inner-filter effect. A: In a dilute solution, the absorption of the excitation light (blue) is small, and the fluorophores (blue spheres) in the center of the cuvette experience nearly the same excitation light intensity (blue) as the fluorophores at the beginning of the light path ($I_{center} \approx I_0$). The intensity of the emitted fluorescence (red) at the center of the cuvette is not attenuated on its path toward the detector ($F_{obs} \approx F_{center}$) B: In highly concentrated solutions, absorption of light by fluorophores at the beginning of the light path reduces the light intensity experienced by fluorophores in the center of the cuvette ($I_{center} < I_0$; primary inner-filter effect). As a consequence, the molecules in the center of the cuvette will be excited less efficiently. Fluorescence emitted from the center of the cuvette may also be partly absorbed ($F_{obs} < F_{center}$; secondary inner-filter effect). The inner-filter effect leads to a breakdown of the linear relationship between fluorescence and concentration of the fluorophore at high fluorophore concentrations.

From the definition of A as the decadic logarithm of I_0/I (see Section 21.2.3), we can calculate the intensity I_{center} at the center from the absorption of the sample at the excitation wavelength, A_{ex}, as

$$I_{center} = I_0 \cdot 10^{-A_{ex}/2} \qquad \text{eq. 21.89}$$

The fluorescence intensity F_{center} is proportional to the intensity I_{center} and hence attenuated by the same factor. F_{center} is additionally reduced by the absorption of the emitted light, A_{em}, on its way from the center to the end of the cuvette, leading to the relationship for the observed fluorescence F_{obs}

$$F_{obs} = F_{center} \cdot 10^{-A_{em}/2} \qquad \text{eq. 21.90}$$

The correct fluorescence intensity F_{corr} in the absence of the inner-filter effect can be calculated by correcting for both parts of the inner-filter effect according to eq. 21.91:

$$F_{corr} = F_{obs} \cdot 10^{(A_{ex}+A_{em})/2} \qquad \text{eq. 21.91}$$

Fluorescent impurities in buffers and other reagents contribute to the overall fluorescence signal and should be avoided by using only the purest chemicals. Reference spectra of the solvent should always be recorded to ensure low fluorescence background. Particles in the solution will scatter the incident light and lead to fluctuations of the readout. Scattering of light also distorts the measured fluorescence spectrum. Rayleigh scattering (Section 23.1.1) occurs at the excitation wavelength. If spectra are measured over a large range of emission wavelengths, the second harmonic of the Rayleigh scattering will be detected at twice the excitation wavelength. For instance, if Trp is

excited at 280 nm, the emission spectrum should be measured only above ~290 nm, and not beyond 2·280 nm = 560 nm to avoid the first and second harmonic of the Rayleigh scattering. Raman scattering (Section 23.1.3) of the excitation light leads to excitation of water vibrations and a scattering peak that might distort the spectrum of the molecule of interest (Box 21.14). Correction of fluorescence spectra for the buffer spectrum (that also contains the Raman peak of water) will eliminate this problem. For time-dependent fluorescence measurements, the contribution from the Raman peak should be eliminated by using suitable cut-off filters.

BOX 21.14: SCATTERING IN FLUORESCENCE SPECTRA.

Rayleigh scattering (Section 23.1.1) of the excitation light superimposes with the fluorescence emission spectrum at the excitation wavelength and twice the excitation wavelength (second harmonic). The intensity for second and higher harmonics is strongly decreased. Because of this and the limited Stokes shift of common fluorophores, higher than second harmonics are typically not detected in fluorescence spectroscopy. Raman scattering (Section 23.1.3) due to the excitation of water vibrations is a common phenomenon in fluorescence spectroscopy of samples in aqueous solutions (Figure 21.57). The position of the Raman peak depends on the excitation wavelength. The wavenumber for the water vibration is 3600 cm^{-1} (Section 21.4). This energy is removed from the incident light, and the scattered light has a higher wavelength (redshift) than the excitation light.

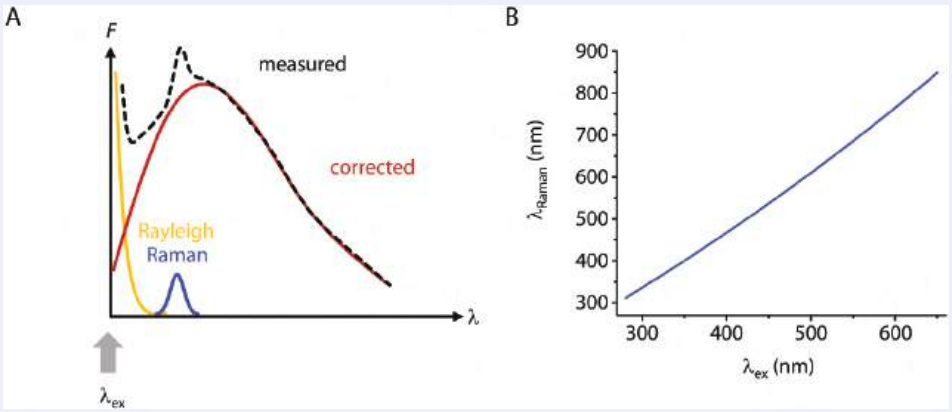


FIGURE 21.57: Scattering in fluorescence emission spectra. A: Rayleigh scattering (orange) is elastic scattering of excitation light. In addition to the first harmonic with $\lambda = \lambda_{ex}$, the second harmonic is observed at $\lambda = 2{\cdot}\lambda_{ex}$ when fluorescence spectra are measured over a wide range of wavelengths. The Stokes band of Raman scattering (blue) is found at wavelengths above λ_{ex}. B: Position of the Raman peak (Stokes band, λ_{Raman}) as a function of the excitation wavelength λ_{ex} (eq. 21.94).

We can calculate the position of the Raman peak in the fluorescence spectrum by converting the excitation wavelength into wavenumbers, subtracting the wavenumber for the water excitation, and re-converting the result to wavelength:

$$\frac{1}{\lambda_{ex}} - 3600\ \text{cm}^{-1} = \frac{1}{\lambda_{Raman}} \qquad \text{eq. 21.92}$$

Solving for the wavelength of Raman scattering, λ_{Raman}, gives

$$\lambda_{Raman} = \frac{\lambda_{ex}}{1 - 3600\ \text{cm}^{-1} \cdot \lambda_{ex}} \qquad \text{eq. 21.93}$$

(*Continued*)

By converting 3600 cm^{-1} into $36 \cdot 10^{-5}$ nm^{-1}, we obtain an equation that gives λ_{Raman} in units of nm when we enter the excitation wavelength λ_{ex} in nm:

$$\lambda_{Raman} = \frac{\lambda_{ex}}{1 - 36 \cdot 10^{-5} \lambda_{ex}} \qquad \text{eq. 21.94}$$

21.5.7 Fluorescence quenching

Fluorescence quenching is a general term for all processes that cause a decrease in fluorescence intensity. Quenching can be static or dynamic, caused by collisions (Figure 21.58). In *collisional quenching*, a fluorophore collides with a quencher molecule during the lifetime of the excited state, and returns to the ground state without emitting a photon. Collisional quenchers provide an alternative process for the depopulation of the excited state, and lead thereby to a decrease in fluorescence lifetime and quantum yield (see Section 21.5.3). In contrast, *static quenching* is caused by formation of a non-fluorescent complex of the quencher and the fluorophore. Intramolecular static quenching occurs in NADH and FAD, where the adenine group stacks on the nicotinamide or flavin moiety (see Section 21.5.4). A static quencher effectively reduces the concentration of the fluorescent species, but does not provide an alternative pathway for relaxation from the excited state to the ground state. Static quenchers therefore do not alter the fluorescence lifetime or the quantum yield.

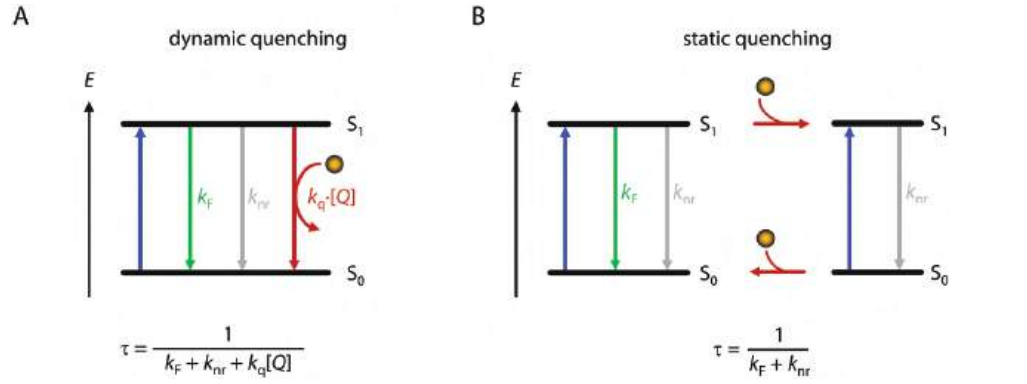


FIGURE 21.58: Collisional and static quenching principle and effect on lifetime. A: Collisional or dynamic quenchers (yellow) quench fluorescence by collisions with the excited state of the molecule and provide an additional path for relaxation of the excited state. The lifetime τ is the inverse sum of the rates for all processes that leads to depopulation of the excited state. Collisional quenching leads to a reduction of the fluorescence lifetime. B: A static quencher binds to the excited state of the fluorophore. The fluorophore-quencher complex relaxes to the ground state by non-radiative processes only. A static quencher reduces the concentration of excited molecules that can undergo fluorescence, but does not provide an extra pathway for depopulation of the S_1 state. As a consequence, the fluorescence lifetime is unaffected by its presence.

One example of a ubiquitous collisional quencher is molecular oxygen. The quenching properties of oxygen are commonly ascribed to its paramagnetic character that allows inter-system crossing to the triplet state. For accurate measurements of fluorescence lifetime and intensity, molecular oxygen dissolved in the samples should be removed. In practice, oxygen levels are often reduced by degassing buffers. On the other hand, the reduction in fluorescence intensity because of collisional quenching by oxygen can be exploited for oxygen sensing. Electron-rich and polarizable compounds such as halide ions (Br^-, I^-) or halogenated compounds and acrylamide also lead to collisional quenching. Trp fluorescence is efficiently quenched by acrylamide.

Collisional quenching can be described quantitatively by the *Stern-Volmer equation*. In the absence of quencher, continuous illumination of the sample leads to a steady-state concentration of fluorophores in the excited state, X*. The differential equation for this steady-state concentration is

$$\frac{\mathrm{d}[X^*]_0}{\mathrm{d}t} = f(t) - (k_\mathrm{f} + k_\mathrm{nr}) \cdot [X^*]_0 = 0 \qquad \text{eq. 21.95}$$

where $f(t)$ is the excitation function that quantifies transitions from the ground state to the excited state under constant illumination, and the second term summarizes the depopulation of the excited state with the rate constant k_f for transitions involving fluorescence, and k_nr for non-radiative deactivation. The index "0" in $[X^*]_0$ refers to the absence of quencher. Note that this differential equation is essentially a rate law and follows the same basic rules (Chapters 7–8; Box 8.2): processes leading to a species (here: a state) enter the balance with a positive sign, processes leading away from this species/state enter with negative signs. Velocities are the products of the respective rate constant and the concentration(s) of the relevant species/state(s).

In the presence of quencher, a third term enters the differential equation to account for the quenching process (eq. 21.96). The velocity of this collisional process is proportional to the quencher concentration $[Q]$ and occurs with a rate constant k_q. We therefore obtain

$$\frac{\mathrm{d}[X^*]}{\mathrm{d}t} = f(t) - (k_\mathrm{f} + k_\mathrm{nr} + k_\mathrm{q}[Q]) \cdot X^* = 0 \qquad \text{eq. 21.96}$$

By equating eq. 21.95 and eq. 21.96, we obtain

$$\frac{[X^*]_0}{[X^*]} = \frac{k_\mathrm{f} + k_\mathrm{nr} + k_\mathrm{q}[Q]}{k_\mathrm{f} + k_\mathrm{nr}} = 1 + \frac{k_\mathrm{q}}{k_\mathrm{f} + k_\mathrm{nr}}[Q] = 1 + K_\mathrm{SV}[Q] \qquad \text{eq. 21.97}$$

The steady-state fluorescence emission F_0 in the absence or F in the presence of quencher is proportional to the concentration of the excited state of the fluorophore ($[X^*]_0$, $[X^*]$), and we can therefore write

$$\frac{F_0}{F} = 1 + K_\mathrm{SV}[Q] \qquad \text{eq. 21.98}$$

which is the *Stern-Volmer equation*. K_SV is the *Stern-Volmer constant* with units M^{-1}. According to the Stern-Volmer equation, the ratio of the fluorescence in the absence and presence of quencher depends linearly on the quencher concentration (Figure 21.59). K_SV is the proportionality constant that can be derived from the slope of the Stern-Volmer plot. It is a measure of the solvent-accessibility of a fluorophore. A high value for K_SV indicates a strong dependence of the fluorescence on the quencher concentration. In this case, the fluorophore is solvent-exposed and accessible to the quencher, which accordingly can very effectively quench the fluorescence. By contrast, a small K_SV indicates low accessibility of the fluorophore.

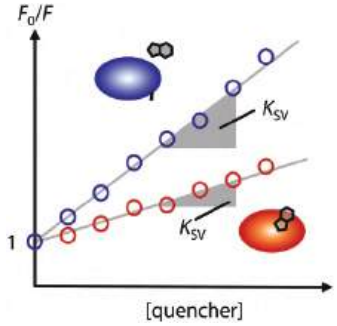


FIGURE 21.59: Stern-Volmer plot. A plot of the ratio F_0/F (fluorescence in the absence of quencher divided by fluorescence at the current quencher concentration) as a function of the quencher concentration is linear (eq. 21.98). From the slope of the plot, the Stern-Volmer constant K_{SV} can be determined. High values for K_{SV}, *i.e.* high sensitivity of the fluorescence on the quencher concentration, are indicative of a solvent-exposed fluorophore (blue: protein with solvent-exposed tryptophan). Low values for K_{SV}, *i.e.* low sensitivity of the fluorescence for the quencher concentration, are indicative of buried fluorophores (red: protein with buried tryptophan).

Fluorescence quenching by small-molecule quenchers is therefore a sensitive probe to determine the solvent accessibility of a fluorophore. For example, surface-exposed tryptophans in proteins will be highly susceptible to quenching, whereas buried tryptophans will not be affected by the addition of quencher (Figure 21.59). In a similar way, quenching can be used to probe the position of a fluorophore in a membrane protein (Figure 21.60). Such an experiment can also provide information on the membrane permeability of the quencher.

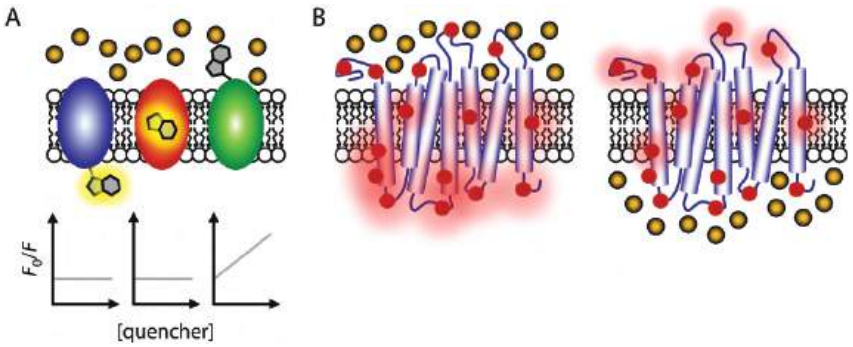


FIGURE 21.60: Fluorescence quenching and the orientation of membrane proteins. A: The exposure of tryptophans (black) toward one side of the membrane can be probed by adding quencher (orange). Fluorescence from tryptophans pointing to the same side will be quenched with increasing quencher concentrations, whereas tryptophans in the interior of the protein and/or the membrane or tryptophans pointing toward the distal compartment do not respond to the addition of quencher (bottom). B: By introducing fluorophores (red) at different positions of the membrane protein and testing their reaction to quencher from both sides of the membrane, the topology of a membrane protein can be determined. The red halo indicates emitting fluorophores, red spheres without halo represent fluorophores whose fluorescence is quenched.

21.5.8 Fluorescence anisotropy

21.5.8.1 Principle of fluorescence anisotropy

So far we have only considered fluorescence emission after excitation of the fluorophores by non-polarized light. When fluorophores are illuminated by linearly polarized light, only those with an orientation that aligns the direction of their transition dipole with the polarization direction of the incident light will be efficiently excited. In solution, the fluorophores have a random orientation, and the linearly polarized light will preferentially excite those that have their transition dipole aligned at

the time of excitation. We have treated this photoselection process in the context of linear dichroism (see Section 21.3.1; Figure 21.22).

What consequence does photoselection have for fluorescence emission? If the fluorophores were to emit fluorescence immediately after excitation, the emitted light would also be polarized in the same plane as the excitation light. However, during the nanosecond lifetime of the excited state, mobile fluorophores have enough time to undergo rotational diffusion (Table 18.12), which eventually randomizes the orientation of the excited fluorophores and their transition dipoles, and the fluorophores emit unpolarized light (Figure 21.61).

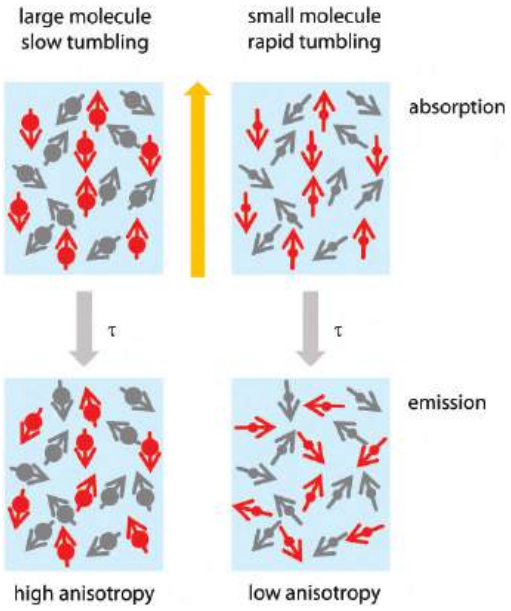


FIGURE 21.61: Photoselection and randomization of the orientation. Excitation with linearly polarized light (orange arrow) preferentially excites those fluorophores whose transition dipoles are aligned with the polarization direction of the incident light (top, red). During the lifetime τ of the excited state, large molecules (left) have rotated very little, and the transition dipoles of the excited fluorophores are still mostly vertically oriented. The polarization direction of the emitted light is therefore the same as that of the incident light. Small molecules (right) rotate much faster, and the orientation of the transition dipoles of the excited fluorophores is randomized during the fluorescence lifetime. Emitted light is therefore not polarized. For simplicity, photoselection is depicted as exclusive excitation of molecules with perfectly aligned transition dipoles.

The velocity of rotational diffusion depends on the size of the rotating moiety. If the rotation is too slow for complete randomization of the fluorophore orientation during the excited state lifetime, the emitted light will not be completely depolarized at the time of emission. Instead, it will contain a component polarized parallel to the polarization plane of the excitation light and a component that is polarized perpendicularly with respect to the excitation light. This phenomenon is called *polarization* or *anisotropy*. In the following, we refer to excitation light that is polarized in the vertical direction with respect to the plane of the instrument table. The component of the emitted light with a polarization parallel to the excitation light is I_v (vertical), the component with perpendicular polarization is I_h (horizontal; Figure 21.62). The polarization p is defined as

$$p = \frac{I_v - I_h}{I_v + I_h} \qquad \text{eq. 21.99}$$

The fluorescence anisotropy r is defined as

$$r = \frac{I_v - I_h}{I_v + 2I_h} \qquad \text{eq. 21.100}$$

Anisotropy and polarization can be interconverted and provide equivalent information. The definition of the anisotropy scales the difference between the two components I_v and I_h of the emitted light by the total intensity ($I_{tot} = I_v + 2I_h$). The horizontally polarized component has to be considered twice because of the two possible orientations perpendicular to the polarization direction of the incident light (Figure 21.62). The anisotropy is therefore independent of the total fluorescence intensity, and additive for mixtures of species with different anisotropies. In contrast, polarization is not additive.

To measure fluorescence anisotropy, a standard steady-state fluorimeter equipped with excitation and emission polarizers is needed (Figure 21.62). The excitation polarizer establishes the vertical polarization direction of the excitation light. The polarizer in the emission path can be aligned in parallel and perpendicular to the excitation polarization to measure I_v and I_h, respectively. From these components, the anisotropy can be calculated according to eq. 21.100.

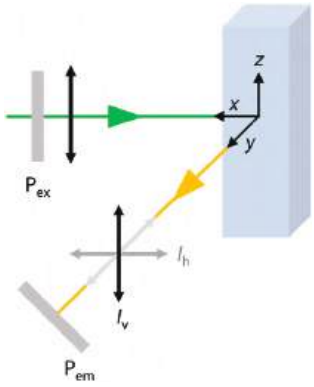


FIGURE 21.62: Measurement of polarization and anisotropy. An excitation polarizer (P_{ex}) selects excitation light polarized along the *z*-axis (green, black double-headed arrow). The components of the emitted light (orange) polarized parallel (I_v, in the *z*-direction) and perpendicular (I_h, in the *x*-direction) to the polarization direction of the incident light are measured by placing an emission polarizer (P_{em}) in front of the detector. Note that the emitted light has a second perpendicular, horizontally polarized component (light gray double-headed arrow), polarized along the *y*-axis, that cannot be measured in this geometry of the instrument. The two horizontally polarized components are identical.

Monochromators and detectors show different sensitivities for vertically and horizontally polarized light. This instrument non-ideality is expressed by the *G-factor*. In the following, we specify the polarization direction of excitation and emission light by two indices. The first letter of the subscript defines the direction of the excitation polarizer, the second letter specifies the direction of the emission polarizer. The G-factor of the instrument is defined as the ratio of the sensitivity for the vertical and parallel components. It is determined by measuring the vertical and horizontal components of the emitted light upon excitation with horizontally polarized light, I_{hv} and I_{hh}, according to

$$G = \frac{I_{hv}}{I_{hh}} \qquad \text{eq. 21.101}$$

The anisotropy r can then be calculated from the measured intensities as:

$$r = \frac{I_{vv} - G \cdot I_{vh}}{I_{vv} + 2G \cdot I_{vh}} \qquad \text{eq. 21.102}$$

21.5.8.2 Applications

Anisotropy is a useful spectroscopic probe to monitor processes in which the tumbling time of the entity carrying the fluorophore changes, typically when a small molecule binds to a macromolecule. The anisotropy increases linearly with degree of binding, therefore an increase in fluorescence anisotropy is frequently used as a probe in equilibrium titration experiments. To maximize the change in anisotropy, the small binding partner (B) is fluorescently labeled, and the binding partner with the high molecular mass (A) is titrated to this small fluorescent ligand (Figure 21.63). The overall tumbling time of the fluorescently labeled partner increases upon complex formation (AB), giving rise to an increase in fluorescence anisotropy. From the corresponding titration curve, the equilibrium dissociation constant of the complex is determined by the same procedure as outlined before (Section 21.5.5.1). When the fluorescence intensity of the fluorescent moiety does not change upon binding, the measured anisotropy r is the population-weighted average of the anisotropy of the free fluorescently labeled molecule, r_{free}, and the bound form, r_{bound}:

$$r = f_{free} \cdot r_{free} + f_{bound} \cdot r_{bound} \qquad \text{eq. 21.103}$$

f_{free} and f_{bound} denote the fractional populations of free and bound forms. The anisotropy can also be expressed in terms of the binding degree, $[AB]/B_t$, as

$$r = r_{free} + \Delta r_{max} \frac{[AB]}{B_t} = r_{free} + \left(r_{bound} - r_{free}\right) \cdot f_{bound} \qquad \text{eq. 21.104}$$

Eq. 21.104 is formally identical to eq. 21.68 that describes the fluorescence signal in a fluorescence (intensity) titration as a function of the degree of binding. Hence, anisotropy titrations of a ligand B with a macromolecule A can be described by an equation analogous to eq. 21.69 to extract the K_d value:

$$r = r_{free} + \Delta r_{max} \frac{\frac{A_t + B_t + K_d}{2} - \sqrt{\left(\frac{A_t + B_t + K_d}{2}\right)^2 - A_t B_t}}{B_t} \qquad \text{eq. 21.105}$$

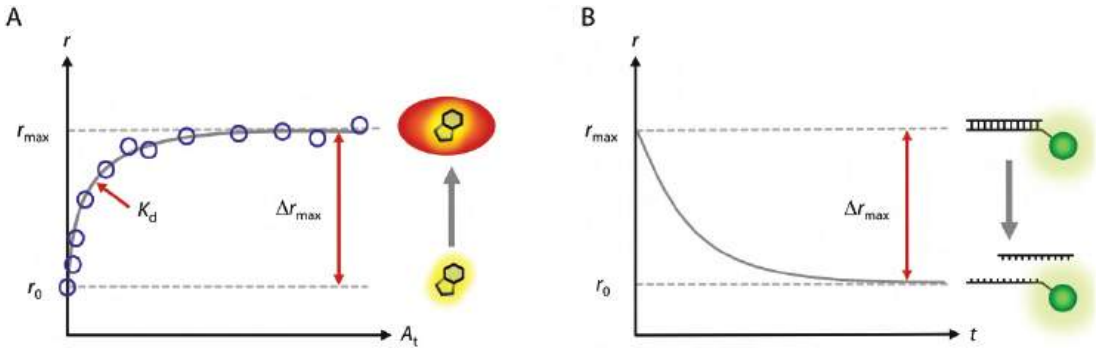


FIGURE 21.63: Fluorescence anisotropy to monitor binding and dissociation. A: A fluorescent ligand, such as a fluorescently labeled nucleotide (yellow) is titrated with a protein (red). The small ligand rotates rapidly, and has a low fluorescence anisotropy (r_0). Upon complex formation with the protein, the mobility of the fluorescent ligand decreases, leading to an increase in anisotropy. At saturation, the anisotropy of the complex, r_{max}, is reached. The dissociation constant K_d of the complex can be determined by analyzing the data points (blue) with eq. 21.105. B: The dissociation of a fluorescently labeled DNA duplex (black/green) can also be monitored *via* fluorescence anisotropy. The fluorophore, attached to one of the strands of the duplex, shows a high fluorescence anisotropy due to the overall slow tumbling of the duplex. Addition of a helicase at $t = 0$ will lead to unwinding of the duplex, and dissociation of the fluorescently labeled strand. The single strand gives a lower fluorescence anisotropy signal because of its more rapid tumbling. The rate constant for duplex unwinding can be determined by describing the observed anisotropy decrease with time by a single exponential.

Note that eq. 21.103 and consequently also eq. 21.105 do not hold for polarization values because the polarization p is not normalized to the total fluorescence intensity.

The steady-state anisotropy of fluorophores generally reports on their mobility on the time scale of the lifetime of the fluorophore (see Section 21.5.9). Membrane-inserted fluorophores such as 1,6-diphenyl-1,3,5-hexatriene (DPH) have been used as probes to study membrane dynamics.

21.5.8.3 Potential pitfalls of polarization/anisotropy measurements

Scattered light is highly polarized, and will lead to overestimation of the actual fluorescence polarization or anisotropy. Scattering should therefore be carefully avoided in anisotropy experiments, or scattered light should be rejected by filters.

Anisotropy and polarization are widely used probes to measure binding. Anisotropy of a mixture of two (or more) species can be calculated as the population-weighted average of the individual anisotropy values because anisotropy values are weighted by the total fluorescence intensity and therefore are additive (eq. 21.103). In contrast, polarization values are not additive, and the polarization of mixtures cannot be calculated as a population-weighted average of the polarization of the components. The deviations are small, but it is important to be aware of this difference when polarization is used as a probe for binding in equilibrium titrations.

An increase in anisotropy that is not caused by scattering is a reliable reporter for binding. However, a lack of change in anisotropy during a titration cannot be taken as a proof that the two molecules do not interact! If the fluorophore is flexibly attached to the one of the partners, the orientation of its transition dipole is largely uncoupled from the orientation of the molecule itself (see Section 21.5.9.2). In this case, the fluorophore has a low anisotropy because of its rapid local motion relative to the molecule both in the free and bound states. A flexibly attached fluorophore therefore does not report on tumbling of the entire molecule, and the anisotropy remains low even if binding occurs. For anisotropy experiments, a rigid attachment of the fluorophore to the molecule of interest is therefore desired. Apart from the local flexibility, the fluorescence lifetime of the fluorophore (see Section 21.5.9.1) also limits its suitability to report on binding. The fluorescence lifetime needs to be on the order of the tumbling times of the free and bound ligand, otherwise the anisotropy is not sensitive to changes in tumbling and molecular mass (see Section 21.5.9.3).

Analysis of anisotropy binding curves for K_d values becomes more complicated when the quantum yield, and thus the fluorescence intensity, changes upon binding. A change in quantum yield means that free and bound forms contribute differently to the measured fluorescence intensities. In this case, the measured overall anisotropy is not only a population-weighted but also an intensity-weighted average. The change in quantum yield has to be taken into account in data analysis (Box 21.15). If the fluorescence intensity changes upon binding, it is advisable to directly use the intensity rather than fluorescence anisotropy as a probe for binding.

BOX 21.15: ANISOTROPY TITRATIONS WITH CHANGES IN QUANTUM YIELD.

Eq. 21.103 for the calculation of the total anisotropy for a mixture of free and bound forms with r_{free} and r_{bound} only holds if both species contribute identically to the fluorescence intensity. If the quantum yield and thus the fluorescence changes upon binding, however, the species with the higher quantum yield contributes more to the overall intensity than the species with the lower quantum yield. For an R-fold change in quantum yield upon binding

$$R = \frac{\phi_{bound}}{\phi_{free}} \qquad \text{eq. 21.106}$$

(*Continued*)

with the quantum yields ϕ_{free} and ϕ_{bound} for the free and bound forms of the fluorescent binding partner, the contribution a_{free} of the free form to the fluorescence intensity is

$$a_{free} = \frac{f_{free}}{f_{free} + R \cdot f_{bound}} \qquad \text{eq. 21.107}$$

f_{free} and f_{bound} refer to the fractional populations of the free and bound forms. The contribution of the bound form is

$$a_{bound} = \frac{R \cdot f_{bound}}{f_{free} + R \cdot f_{bound}} \qquad \text{eq. 21.108}$$

These relative contributions a_{free} and a_{bound} are the weighting factors for the calculation of the overall anisotropy r from the individual anisotropy values r_{free} and r_{bound}:

$$r = \frac{f_{free}}{f_{free} + R \cdot f_{bound}} \cdot r_{free} + \frac{R \cdot f_{bound}}{f_{free} + R \cdot f_{bound}} \cdot r_{bound} \qquad \text{eq. 21.109}$$

We can now substitute f_{free} by $1–f_{bound}$. For a binary equilibrium of A and B with the AB complex, f_{bound} can be expressed in terms of A_t, B_t, and K_d by the solution of the quadratic equation (eq. 21.65 and eq. 21.66). We thereby obtain an equation for the anisotropy r as a function of A_t, B_t, K_d, r_{free}, r_{bound}, and R that we can use to describe an experimental binding curve. The increase in quantum yield R is typically determined from reference spectra of the free and bound species, and kept constant during the fit.

21.5.9 Time-resolved fluorescence

In the previous Sections (Sections 21.5.1–21.5.8), we have discussed fluorescence emission during steady-state illumination of the sample. What happens if we illuminate a fluorophore only briefly, with a short light pulse, and then monitor the emitted fluorescence? During illumination, the excited state of the fluorophore, X*, becomes populated. The fluorophores in the excited state will then relax to the ground state. The fluorescence $F(t)$ at any given time is proportional to the concentration of molecules in the excited state $[X^*]$

$$F(t) \propto [X^*] \qquad \text{eq. 21.110}$$

The velocity of the relaxation depends on the rate constant k for all processes leading away from the excited state

$$\frac{d[X^*]}{dt} = k \cdot [X^*] \qquad \text{eq. 21.111}$$

where k is the sum of the rate constants for radiative and non-radiative deactivation processes. The inverse of k is the fluorescence lifetime τ, the mean time a fluorophore spends in the excited state before returning to the ground state (see Section 21.5.3). The relaxation, accompanied by the emission of fluorescence, will depopulate the excited state and the fluorescence will decay to zero. By separating the variables in eq. 21.111 and integrating over the time t, we obtain an expression for the concentration of the fluorophore in the excited state as a function of time:

$$[X^*](t) = [X^*]_0 \cdot \exp(-kt) \qquad \text{eq. 21.112}$$

Eq. 21.112 can be combined with eq. 21.110 to arrive at the exponential function that describes the *fluorescence decay* $F(t)$:

$$F(t) = F_0 \cdot \exp(-kt) = F_0 \cdot \exp\left(-\frac{t}{\tau}\right) \qquad \text{eq. 21.113}$$

F_0 is the fluorescence intensity at time $t = 0$, and τ is the *fluorescence lifetime*, the time at which the fluorescence intensity has decayed to F_0/e. If different fluorescent species are present, the fluorescence decay is described by a sum of exponentials with individual lifetimes τ_i and corresponding amplitudes α_i:

$$F(t) = \sum_i \alpha_i \cdot e^{-t/\tau_i} \qquad \text{eq. 21.114}$$

The amplitude values α_i correspond to the fraction of the species i in the mixture (Figure 21.64). Multiple fluorescence lifetimes are observed for mixtures of fluorophores with different fluorescence lifetimes. One type of fluorophore can also give rise to more than one lifetime if it can exist in different chemical environments (free or bound, solvent-exposed or buried). In this case, fluorescence lifetime measurements provide information on the relative abundance of the different species, which is proportional to the amplitude α_i of the lifetime component τ_i. In fluorescence intensity measurements under steady-state illumination, this information is averaged out.

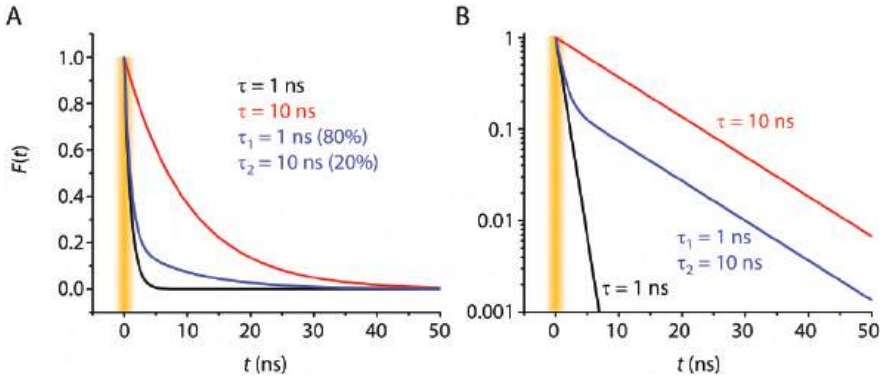


FIGURE 21.64: Fluorescence decay. A: When fluorophores are excited by a short light pulse at $t = 0$ (orange), their fluorescence decays exponentially. After the lifetime τ, the fluorescence has decreased to $1/e$ of its initial value. Fluorophore A has a lifetime of 1 ns (black), fluorophore B of 10 ns (red). A mixture of 80% A and 20% B shows a biexponential decay (blue). Biexponential decays are also obtained for the same fluorophore in different environments (*e.g.* free and bound). B: Same fluorescence decays as in A plotted with a logarithmic *y*-axis.

21.5.9.1 Measurement of fluorescence lifetimes

Fluorescence lifetimes are measured by exciting fluorophores with a very short laser pulse, typically a few to a few hundred ps, and then following the decay as a function of time. For technical reasons, the fluorescence cannot be measured directly on the nanosecond time scale. Therefore, a technique called *time-correlated single photon counting* is used (Figure 21.65). Time-correlated single photon counting measures the arrival time of the first photon emitted from the sample after the excitation pulse. The arrival times of photons measured in a large number of excitation-emission cycles are histogrammed. From the envelope of these arrival time histograms, the fluorescence decay is obtained.

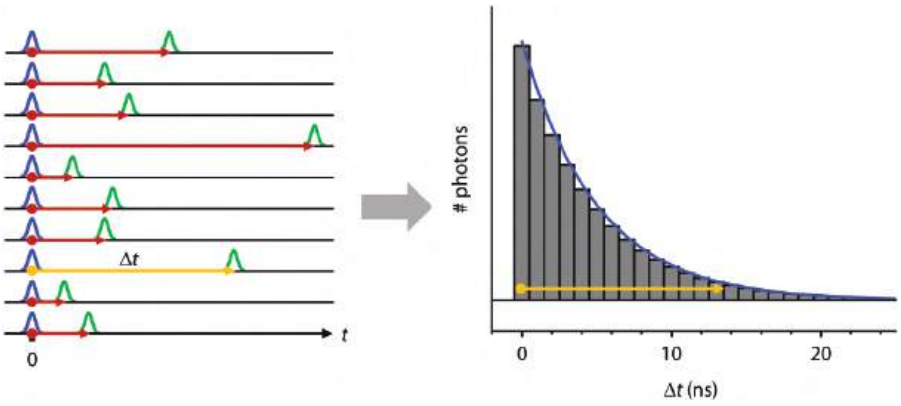


FIGURE 21.65: Measurement of fluorescence lifetimes by time-correlated single photon counting: From histograms to decay curves. Fluorescence is excited by a short laser pulse (blue) at $t = 0$, and the first photon emitted from the sample is detected (green). The time between excitation and emission (red arrow) is measured. The measurement is repeated, and the determined time differences between excitation pulse and emission events are histogrammed (right). The envelope of such a histogram is the fluorescence decay (red). The orange arrow indicates the time difference Δt for one single excitation/emission cycle (left), and its position in the histogram (right).

Strictly speaking, time-correlated single photon counting yields fluorescence decays and lifetimes directly only when the excitation pulse is much shorter than the fluorescent lifetime. However, this is not always the case. Due to the finite length of the excitation pulse, a series of fluorescence decays is initiated over the duration of the excitation pulse (Figure 21.66). The amplitude of each individual decay is proportional to the intensity of the excitation light at that time point. The measured fluorescence decay is the sum of all of these individual decays, with different amplitudes and starting times. Mathematically, the measured decay is a *convolution* (see Section 31.16) of the *instrument response* (or lamp function) and the true fluorescence decay (Box 21.16). The instrument response corresponds to the excitation pulse filtered by the detection electronics. It is measured in a reference measurement with a scattering solution as a sample. The true fluorescence decay can then be retrieved from the fluorescence decays measured by time-correlated single photon counting by deconvolution with the instrument response function.

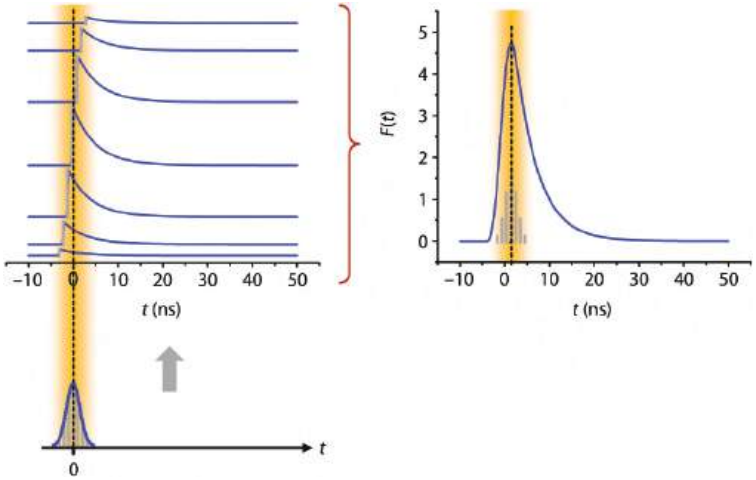


FIGURE 21.66: Instrument response and measured decay. The excitation pulse used in time-resolved fluorescence measurements is not an ideal, infinitely short pulse at $t = 0$ but can often be described by a Gaussian intensity distribution, centered at $t = 0$ (bottom left, blue). The exact shape of this lamp function depends on the excitation source used. The excitation pulse consists of a number of infinitely short pulses (gray bars). Each of these individual pulses elicits an impulse response in the sample, *i.e.* a fluorescence decay with the amplitude of the excitation light intensity at this point in time. Therefore, a finite pulse generates a series of instrument responses with different amplitudes and different starting times (top left). The sum of these individual responses is the measured fluorescence intensity (right). Mathematically, the measured fluorescence intensity is the convolution of the lamp function and the true fluorescence decay (Box 21.16).

BOX 21.16: CONVOLUTION AND DECONVOLUTION.

The measured fluorescence intensity in lifetime experiments is the convolution of the lamp function and the true fluorescence decay. The lamp function can be described as a series of δ-functions. A δ-function is defined as a function that has a certain value for one x-value, but has a value of zero for all other x-values. To calculate the impulse response $I_k(t)$ to each δ-function, we have to shift the true fluorescence decay $I(t)$ along the time axis by t_k. The shifted decay $I(t–t_k)$ is then multiplied by the height $L(t_k)$ of this δ-function and by the time difference Δt to the next δ-function:

$$I_k(t) = L(t_k) \cdot I(t - t_k)\Delta t \qquad \text{eq. 21.115}$$

The measured fluorescence intensity $F(t_k)$ is the sum of all these individual impulse responses $I_k(t)$:

$$F(t_k) = \sum_{t=0}^{t_k} L(t_k) \cdot I(t - t_k)\Delta t \qquad \text{eq. 21.116}$$

For infinitely small intervals, *i.e.* small Δt values, we can replace the sum by an integral and integrate from $t = 0$ to t' (corresponding to t_k):

$$F(t) = \int_0^{t'} L(t') \cdot I(t - t')\mathrm{d}t' \qquad \text{eq. 21.117}$$

By replacing t' with t-μ, we obtain

$$F(t) = \int_0^{t} L(t - \mu) \cdot I(\mu)\mathrm{d}\mu \qquad \text{eq. 21.118}$$

Eq. 21.118 is the mathematical definition of the convolution of two functions $L(t$-μ$)$ and $I(t)$, also written as

$$F(t) = L(t) \otimes I(t) \qquad \text{eq. 21.119}$$

Convoluting two functions with each other means shifting the mirror image of one (here $L(t)$) with respect to the y-axis relative to the other (here $I(t)$) by μ, and calculating the product. We will see later that this definition is different from the definition of the auto- and cross-correlation (Section 23.1.2, eq. 23.17; Sections 25.6.1 and 25.6.3, eq. 25.23 and eq. 25.32), where the first function is not inverted (Figure 21.67). We will revisit the convolution in X-ray crystallography, where a macromolecular crystal can be described as the convolution of a lattice function and the function for the entity that is found at each lattice point (Section 24.4.2).

(*Continued*)

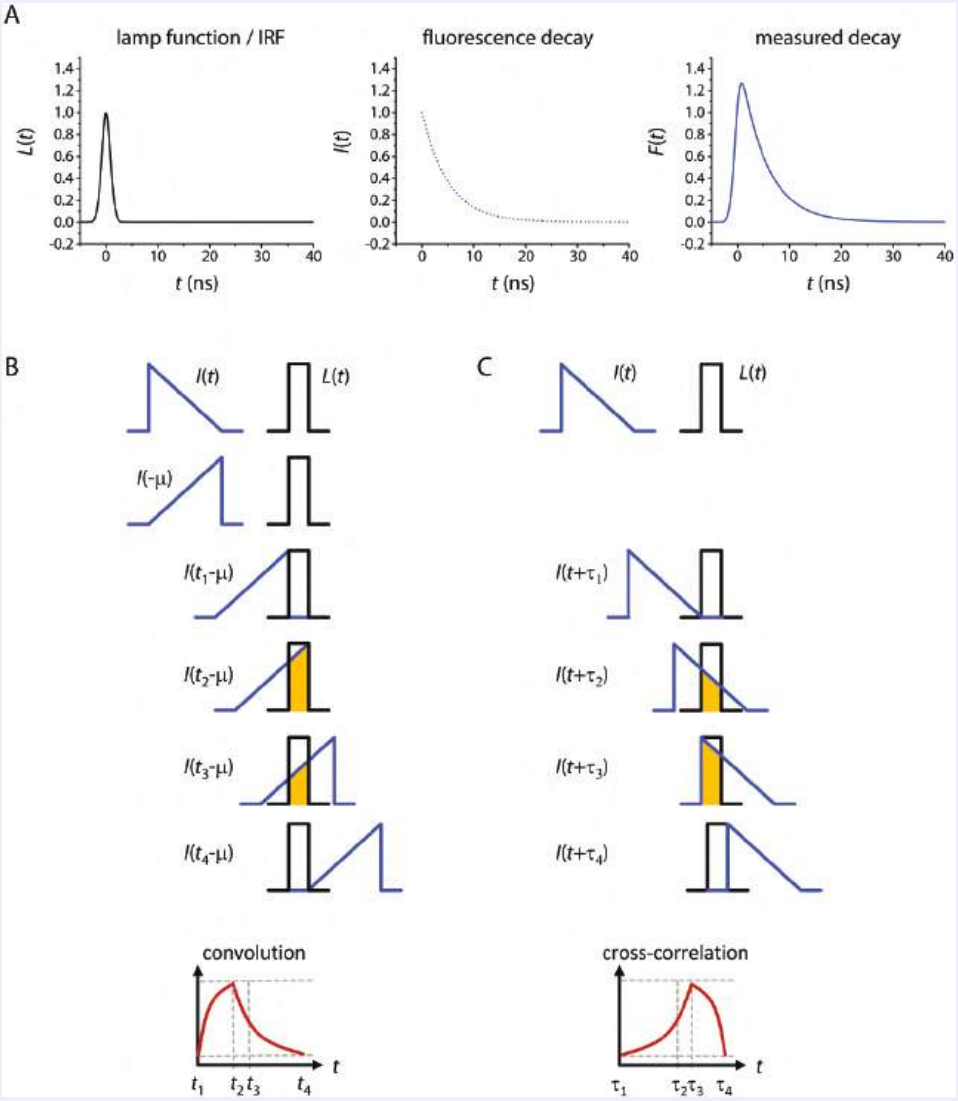


FIGURE 21.67: Convolution. A: The measured fluorescence decay $F(t)$ is a convolution of the lamp function $L(t)$, also called instrument response function, and the true fluorescence decay $I(t)$. B: The convolution of two functions can be constructed by inverting one function, shifting it relative to the second function, and calculating the area under the two curves. This is illustrated for the convolution of $I(t)\otimes L(t)$: the fluorescence decay ($I(t)$, blue) is inverted to $I(-\mu)$ and shifted relative to the lamp function ($L(t)$, black). The area under both curves is highlighted in orange for different time points t_1-t_4. The convolution $I(t)\otimes L(t)$ (red) is identical to the convolution $L(t)\otimes I(t)$. C: In comparison, auto- and cross-correlation are calculated by shifting one function relative to itself (auto-correlation) or to the second function (cross-correlation), and calculating the area under the curves. Here, the cross-correlation of $L(t)$ and $I(t)$ is illustrated (red). Note that $I(t)$ is not inverted to calculate the cross-correlation. Auto- and cross-correlation are treated in Section 25.6.

21.5.9.2 Fluorescence anisotropy decays and rotational correlation times

Instead of measuring the decay of the (isotropic) fluorescence intensity, fluorophores can also be excited by a pulse of linearly polarized light. From the time-dependent fluorescence emission with a polarization parallel and perpendicular to the polarization of the excitation light, the decay $r(t)$ of the fluorescence anisotropy can be calculated from the general definition of anisotropy (Section 21.5.8) as

$$r(t) = \frac{I_v(t) - I_h(t)}{I_v(t) + 2I_h(t)} \qquad \text{eq. 21.120}$$

Note that eq. 21.120 refers to excitation with light polarized in the vertical direction with respect to the plane of the instrument table, and $I_v(t)$ is the parallel component, $I_h(t)$ the perpendicular component. Anisotropy decays can then be described as sums of exponential functions, with amplitude factors β_i and time constants that are now called *rotational correlation time* $\tau_{c,i}$.

$$r(t) = \sum_i \beta_i \cdot e^{-t/\tau_{c,i}} \qquad \text{eq. 21.121}$$

The fluorescence anisotropy decays of fluorescently labeled (bio-)macromolecules can often be described by a sum of two exponentials, with a high rotational correlation time corresponding to the overall tumbling of the fluorophore with the macromolecule, and a small correlation time that reflects the *segmental mobility* of the fluorophore relative to the macromolecule (Figure 21.68). The segmental mobility can be approximated by the "wobbling in a cone" model, in which the linker connecting the fluorophore to the macromolecule can access the inside of a cone. From the amplitude of the two phases in the anisotropy decay, the order parameter S can be calculated according to eq. 21.122:

$$S = \sqrt{\frac{\beta_2}{\beta_1 + \beta_2}} \qquad \text{eq. 21.122}$$

S is related to the half-cone angle θ of the cone within which the fluorophore can move by eq. 21.123:

$$\theta = \arccos\left(\frac{1}{2}\left(\sqrt{1+8S} - 1\right)\right) \qquad \text{eq. 21.123}$$

The larger the half-cone angle, the larger is the conformational space accessible to the fluorophore. By determining the order parameter and half-cone angle from fluorescence anisotropy decays, the rigidity of a fluorophore-macromolecule linkage can be probed.

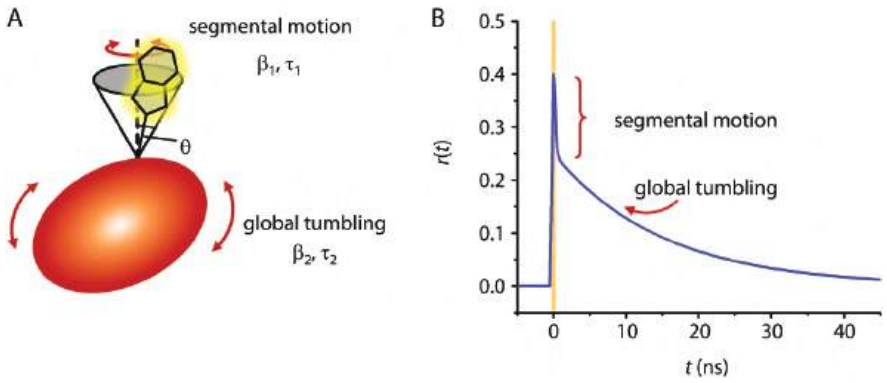


FIGURE 21.68: "Wobbling in a cone" model: segmental flexibility gives rise to two rotational correlation times. A: A fluorophore (yellow) attached to a macromolecule (orange) by a linker can move within a cone (black) because of the segmental flexibility relative to the macromolecule. B: The anisotropy decay (blue) of this fluorophore is described by two rotational correlation times: a small rotational correlation time τ_1 accounts for the rapid segmental motion of the fluorophore relative to the macromolecule. A larger rotational correlation time τ_2 represents the global tumbling of the fluorophore with the macromolecule. From the amplitudes β_1 and β_2 of the two components of the anisotropy decay, the half-cone angle θ can be calculated (eq. 21.123).

21.5.9.3 Rotational correlation time and molecular size

The rotational correlation time depends on the size of the fluorescent molecule and the viscosity of the medium. Simply speaking, a small molecule will rotate rapidly, and show a small rotational correlation time, *i.e.* the anisotropy decays rapidly. For large molecules, tumbling is slow, the rotational correlation time is large, and the anisotropy will decay slowly. High viscosities lead to slow tumbling, low viscosities allow rapid tumbling.

Similar to translational motion (see Sections 4.1, 23.1.2), rotational motion can be described by a rotational diffusion coefficient D_{rot}. The diffusion coefficient for rotational motion depends on the rotational friction coefficient f_{rot} of the molecule according to eq. 21.124:

$$D_{rot} = \frac{k_B T}{f_{rot}} = \frac{k_B T}{6 V_h \eta} \qquad \text{eq. 21.124}$$

with the Boltzmann constant k_B, the temperature T, the hydrated molecular volume V_h, and the viscosity η. The rotational correlation time τ_c is defined as

$$\tau_c = \frac{1}{6 D_{rot}} = \frac{V_h \eta}{k_B T} \qquad \text{eq. 21.125}$$

τ_c thus depends linearly on the ratio η/T. Measurements of τ_c as a function of viscosity at constant temperature will give a linear plot with the hydrated molecular volume V_h as the slope. The specific hydrated volume for proteins is typically 1 $cm^3\ g^{-1}$. To convert the specific hydrated volume into the hydrated volume for a single molecule, we need to multiply by the molecular mass and divide by the Avogadro constant N_A. From eq. 21.125, we can then estimate a rotational correlation time of 1 ns per 2800 Da molecular mass at 25°C. This rule of thumb only holds for spherical molecules.

The steady-state anisotropy (under constant illumination) can be calculated from the fluorescence decay $I(t)$ and the anisotropy decay $r(t)$ according to

$$r = \frac{\int_0^\infty I(t) r(t) \mathrm{d}t}{\int_0^\infty I(t) \mathrm{d}t} \qquad \text{eq. 21.126}$$

For a spherical molecule, the anisotropy decay will follow a single exponential:

$$r(t) = r_0 \cdot e^{-t/\tau_c} \qquad \text{eq. 21.127}$$

Assuming that the intensity decay also follows a single exponential, we can express $r(t)$ and $I(t)$ as exponential functions and evaluate the integral, which gives

$$r = \frac{r_0}{1 + \frac{\tau}{\tau_c}} \qquad \text{eq. 21.128}$$

According to eq. 21.128, the anisotropy r depends on the ratio of the fluorescence lifetime τ and the rotational correlation time τ_c. In the extreme case of a static molecule, with $\tau_c \to \infty$, the anisotropy is $r = r_0$. This is also true for $\tau \ll \tau_c$. For $\tau \gg \tau_c$, on the other hand, the anisotropy r will be very low and almost independent of the tumbling time τ_c. A fluorophore thus reports on changes of the tumbling time only when its lifetime is on the order of the rotational correlation time of the species studied (see Section 21.5.8.3).

Eq. 21.128 can be rearranged to

$$\frac{1}{r} = \frac{1}{r_0} + \frac{\tau}{r_0 \tau_c} \qquad \text{eq. 21.129}$$

This equation is called the *Perrin equation*. Using the relationship in eq. 21.125, we can also express the Perrin equation as

$$\frac{1}{r} = \frac{1}{r_0} + \frac{\tau k_B T}{r_0 V_h \eta} \qquad \text{eq. 21.130}$$

According to this equation, a plot of $1/r$ as a function of T/η (or τ) is linear (Figure 21.69). From the intercept, r_0 can be determined. The slope of the plot yields V_h, the hydrated volume of the molecule of interest. For fluorescently labeled biomolecules, segmental mobility of the chromophore often causes an increase of the observed $1/r$, *i.e.* a decrease in the anisotropy r. The extrapolated r_0 is then also smaller than the true r_0. In fluorescence anisotropy decays, this segmental motion leads to a second, smaller rotational correlation time in anisotropy (Figure 21.68).

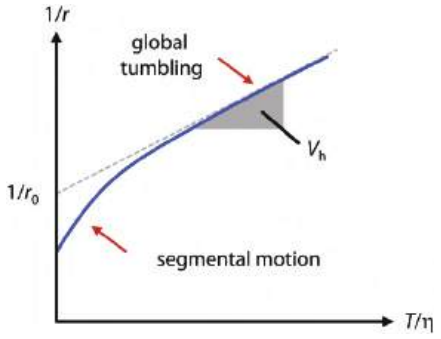


FIGURE 21.69: Perrin plot for macromolecules and chromophores. A plot of $1/r$ as a function of T/η is linear (eq. 21.129). From the slope, the hydrated volume V_h of the molecule can be determined (eq. 21.130). If a fluorophore is flexibly attached to a macromolecule, its segmental motion will cause a deviation from linearity at low temperatures/high viscosities, and the limiting anisotropy extrapolated from the linear region of the plot (gray broken line) is lower than the true value.

21.5.9.4 Applications

Fluorescence lifetimes report on the local environment of the fluorophore. 2-aminopurine is a fluorescent nucleobase analog whose lifetimes are a sensitive probe for nucleic acid structure and dynamics. In solution, 2-AP fluorescence decays with a single lifetime of about 10 ns. 2-AP fluorescence is quenched within a DNA duplex due to hydrogen bonding and because of stacking with neighboring bases. Quenching is alleviated when stacking or base pairing are perturbed. When 2-AP is incorporated into double-stranded DNA, its fluorescence decay is complex, and multiple lifetimes are observed. Although these lifetimes are difficult to assign to a particular structural change, changes in observed lifetimes and amplitudes are a sensitive qualitative probe for changes in the local environment. Fluorescence lifetimes are also measured in time-resolved FRET experiments to resolve multiple species with different donor-acceptor distances (see Section 21.5.10).

Fluorescence lifetimes are also used for fluorescence sensing. A fluorescence lifetime sensor provides a concentration-independent read-out of changes in steady-state intensities and fluorescence quenching. An example is the fluorescein-coupled chelator Calcium Green, whose fluorescence lifetime increases upon Ca^{2+} binding. Calcium Green is used in live-cell fluorescence lifetime imaging (see Section 25.5).

Measurement of rotational correlation times can provide unambiguous evidence for binding, even in cases where no increase in steady-state anisotropy is observed. A large segmental flexibility of the fluorophore leads to a low steady-state fluorescence, and may preclude changes in steady-state anisotropy upon binding to a higher molecular mass partner. Anisotropy decays, on the other hand, directly provide rotational correlation times for segmental flexibility and for global tumbling and the associated amplitudes. An increase in the rotational correlation time associated with global tumbling in the presence of a potential binding partner is clear evidence for binding. Rotational correlation times are also measured to quantify the local flexibility of fluorophores attached to

biomolecules (eq. 21.122, eq. 21.123), for example as controls in Förster resonance energy transfer experiments (Section 21.5.10).

21.5.10 Förster Resonance Energy Transfer

A single fluorophore absorbs excitation light and then emits fluorescence at its characteristic wavelength and with a characteristic fluorescence lifetime. If a second fluorophore is present in the vicinity, the transition dipoles are not independent of each other, but interact in a distance-dependent manner (see Section 17.2.2.2). Due to this dipolar coupling, the first fluorophore can transfer its excitation energy to the second fluorophore. As a result, the first fluorophore will return to the ground state in a radiation-less process, and the second fluorophore undergoes a transition to the excited state. Upon relaxation to the ground state, the second fluorophore will emit fluorescent light. Because the energy is donated by the first fluorophore, it is called the *donor*. The second fluorophore taking up the energy is called the *acceptor*. The whole process of radiation-less energy transfer from a donor to an acceptor fluorophore is called *Förster resonance energy transfer* (FRET), after Theodor Förster who described the process quantitatively in 1948.

21.5.10.1 Principle of FRET

As a dipolar interaction, the efficiency of FRET depends on the inverse 6th power of the distance (see van der Waals interactions in Section 17.2.2.4 and the nuclear Overhauser effect in NMR spectroscopy, Section 22.1.3). Because of its strong distance dependence, FRET can be used as a molecular ruler on the relevant length scales for biomolecules. The rate constant k_t for FRET depends on the inter-fluorophore distance r according to

$$k_t = \frac{1}{\tau_D} \cdot \left(\frac{R_0}{r} \right)^6 \qquad \text{eq. 21.131}$$

In eq. 21.131, R_0 is the *Förster distance* or *characteristic transfer distance*. R_0 depends on the spectral properties of the two fluorophores between which FRET occurs (see eq. 21.135).

The efficiency E of the energy transfer from the donor to the acceptor is determined by the kinetic competition (Section 8.2; Figure 21.70) between energy transfer (k_t) and all other processes leading to a depopulation of the excited state, including donor fluorescence ($k_{D,f}$) or non-radiative donor relaxation ($k_{D,nr}$)

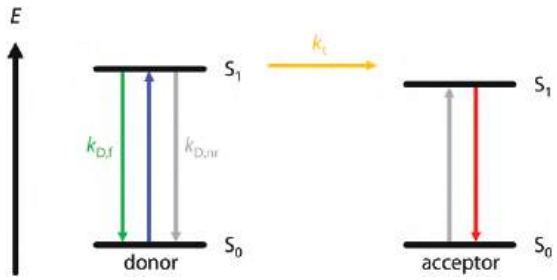


FIGURE 21.70: Jablonski diagram for FRET. The donor fluorophore is excited (blue arrow). From the excited state S_1, it can either relax to the ground state by radiative ($k_{D,f}$, green) or non-radiative ($k_{D,nr}$, gray) transitions, or transfer its energy to a suitable acceptor. Due to this non-radiative energy transfer (k_t, orange), the donor relaxes back to the ground state S_0 without fluorescence emission, and the acceptor transitions to the excited state. Once the acceptor relaxes to the ground state, its fluorescence can be observed (red arrow).

$$E = \frac{k_t}{k_{D,f} + k_{D,nr} + k_t} \qquad \text{eq. 21.132}$$

We can express the rate constants for relaxation of the donor *via* radiative and non-radiative processes as the inverse of the donor fluorescence lifetime τ_D as

$$\frac{1}{\tau_D} = k_{D,f} + k_{D,nr} \qquad \text{eq. 21.133}$$

By combining eq. 21.131 through eq. 21.133, we obtain the dimensionless FRET efficiency as

$$E = \frac{R_0^6}{R_0^6 + r^6} \qquad \text{eq. 21.134}$$

with its dependence on the inverse 6th power of the distance r separating the dyes. Due to its prominent distance dependence, the FRET efficiency can be used as a molecular ruler. The Förster distance R_0 is the distance at which the energy transfer efficiency is 0.5 or 50%. It is defined by the spectral properties of donor and acceptor fluorophores, and can be calculated according to the relationship

$$R_0^6 = 8.785 \cdot 10^{-11} \frac{\kappa^2 \phi_D J}{n^4} \qquad \text{eq. 21.135}$$

In eq. 21.135, κ^2 is the orientation factor, ϕ_D is the quantum yield of the donor, J is the overlap integral and n is the refractive index of the medium. Both J and ϕ_D can be determined experimentally. The quantum yield ϕ_D of the donor can be determined in comparison to reference fluorophores (see Box 21.8). The *overlap integral J* is a measure of the spectral overlap of the fluorescence emission of the donor and the absorption of the acceptor. It can be determined from the fluorescence emission spectrum of the donor, $F_D(\lambda_i)$ normalized to a total intensity of 1, and the absorption spectrum of the acceptor, normalized to the molar extinction coefficient $\varepsilon_A(\lambda_i)$, according to eq. 21.136:

$$J = \int_0^\infty F_D(\lambda) \cdot \varepsilon_A(\lambda) \cdot \lambda^4 \, d\lambda \approx \sum_i F_D(\lambda_i) \cdot \varepsilon_A(\lambda_i) \cdot \lambda_i^4 \qquad \text{eq. 21.136}$$

Typically, the two spectra are measured with a step size of 1 nm, and the overlap integral is then approximated by the sum in eq. 21.136 (Figure 21.71).

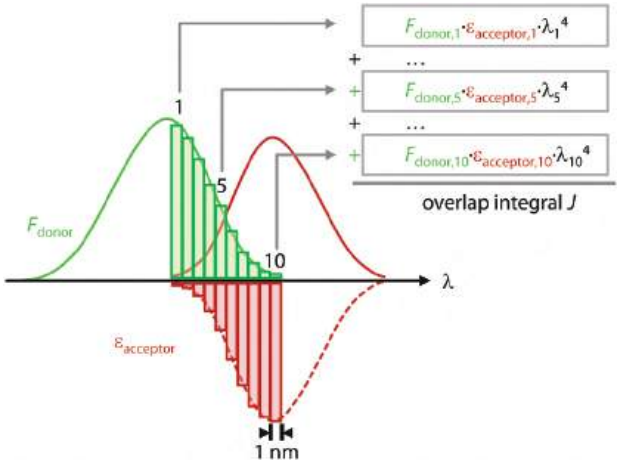


FIGURE 21.71: Calculation of the overlap integral. The overlap integral J can be calculated from the fluorescence emission spectrum of the donor, normalized to an area under the emission peak of 1 (green) and the absorption spectrum of the acceptor in terms of the extinction coefficient ε (red, dotted red line: inverted spectrum for better visibility). The fluorescence and absorption values, measured in 1 nm steps (green and red bars) are multiplied for each wavelength λ in the overlapping region, and each value is weighted by λ^4, shown exemplarily for the values 1, 5, and 10. The sum of all values is the overlap integral J (eq. 21.136).

The *orientation factor* κ^2 is a measure of the relative orientation of the transition dipoles of donor and acceptor (Figure 21.72). It depends on the angles θ_D and θ_A of the donor and acceptor transition dipoles with the vector joining them, and the angle θ_{DA} between donor and acceptor transition dipoles. The angle θ_{DA} is the dihedral angle of the two planes defined by the joining vector and the donor and acceptor transition dipole, respectively.

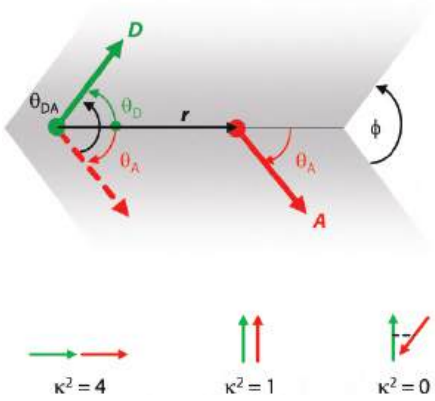


FIGURE 21.72: The orientation factor κ^2. The orientation factor κ^2 describes the relative orientation of the transition dipoles of donor and acceptor fluorophores (eq. 21.137). It depends on the angle θ_D of the donor transition dipole (green) with the vector ***r*** (black) joining the transition dipoles of donor (green) and acceptor (red), the angle θ_A of the acceptor transition dipole (red) with the vector ***r***, and the angle θ_{DA} between donor and acceptor transition dipoles. This angle θ_{DA} is the dihedral angle ϕ of the two planes defined by the vectors ***r*** and the donor and acceptor transition dipole, respectively.

$$\kappa^2 = \left(\cos\theta_{DA} - 3\cos\theta_D \cdot \cos\theta_A\right)^2 \qquad \text{eq. 21.137}$$

κ^2 can cover a range of values from zero to four. Clearly, its value has to be known if changes in FRET efficiency are to be interpreted in terms of distance changes. Unfortunately, transition dipole directions are not straightforward to determine, and often unknown. κ^2 adopts a value of ⅔ when both donor and acceptor fluorophore can sample all possible relative orientations on the time scale of the experiments. To ensure that this condition is met, fluorophores for FRET are ideally attached to the molecule of interest by flexible linkers, such that they can almost freely rotate relative to the molecule of interest. A flexible attachment of fluorophores also minimizes interactions with the labeled biomolecule that may lead to fluorescence changes, and prevents interference with its function. The segmental mobility of fluorophores attached to a macromolecule can be estimated from fluorescence anisotropy decay parameters (see Section 21.5.9).

21.5.10.2 Experimental determination of FRET efficiencies

The *FRET efficiency E* from the donor can be determined experimentally from the donor and acceptor fluorescence intensities F_D and F_A that are measured after excitation of the donor fluorophore (eq. 21.138):

$$E = \frac{F_A}{F_D + F_A} \qquad \text{eq. 21.138}$$

The donor and acceptor contributions can be measured at fixed wavelengths for donor and acceptor emission. More accurately, they are determined from steady-state emission spectra $F_{DA}(\lambda)$ of the molecule that carries both fluorophores (Figure 21.73). Such a spectrum is a linear combination, or weighted sum, of the emission spectra of the donor $F_D(\lambda)$ and of the acceptor $F_A(\lambda)$. These reference spectra are obtained by measuring the fluorescence emission spectrum of the molecule of interest that is labeled with only the donor or only the acceptor.

$$F_{DA}(\lambda) = f_D \cdot F_D(\lambda) + f_A \cdot F_A(\lambda) \qquad \text{eq. 21.139}$$

From the weighting coefficients f_D and f_A, the FRET efficiency can be calculated (eq. 21.140):

$$E = \frac{f_A}{f_D + f_A} \qquad \text{eq. 21.140}$$

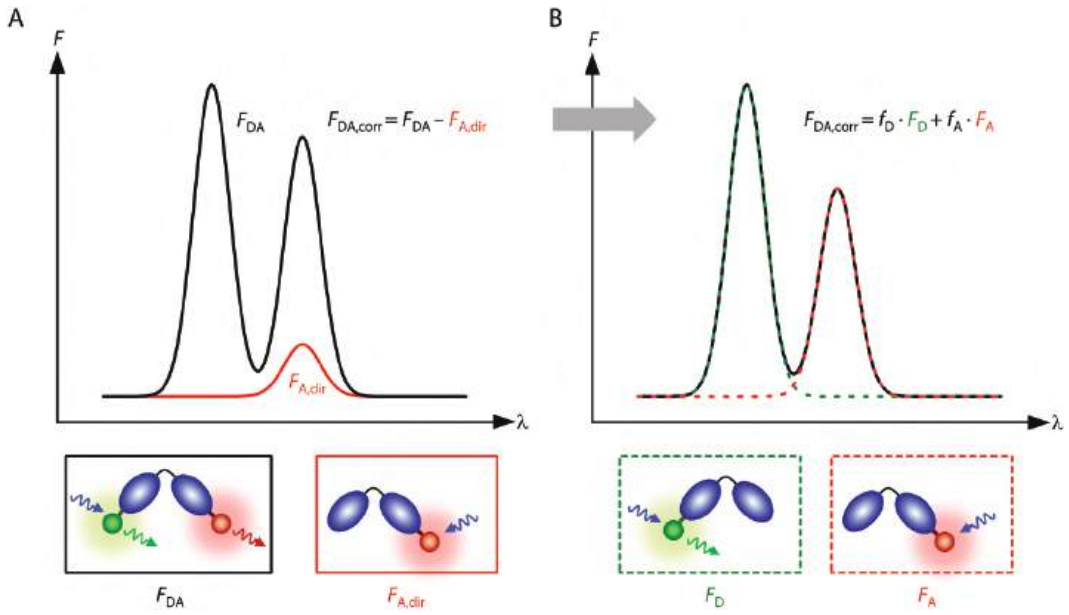


FIGURE 21.73: Determination of the FRET efficiency from donor and acceptor intensities. A: The fluorescence emission spectrum of the donor/acceptor-labeled molecule upon donor excitation is recorded (F_{DA}, black) and corrected for the contributions from the direct excitation of the acceptor at the donor excitation wavelength ($F_{A,dir}$, red). B: The corrected spectrum $F_{DA,corr}$, calculated by subtracting the measured spectra in A, is described as a weighted sum (linear combination) of the donor fluorescence (F_D, green) and the acceptor fluorescence (F_A, red). The weighting coefficients f_D and f_A are used to calculate the FRET efficiency E using eq. 21.140.

Methods that calculate the FRET efficiency from measurements of donor and acceptor fluorescence are called *ratiometric methods*. Measuring both donor and acceptor fluorescence provides the inherent control of anti-correlated changes in donor and acceptor fluorescence when FRET occurs. Any change in donor or acceptor fluorescence without the corresponding opposite change in fluorescence of the second dye can be attributed to other events than FRET, such as donor or acceptor quenching. Ratiometric approaches are also not affected by errors in concentration of the donor/acceptor-labeled molecule, or by instabilities of the excitation source. Alternatively, the FRET efficiency can be determined by measuring either the effect of the acceptor on the donor fluorescence (donor quenching) or the effect of the presence of the donor on the acceptor fluorescence (acceptor sensitization). To calculate the FRET efficiency from donor quenching, the donor fluorescence in the absence of acceptor, F_D, is determined in a fluorescence measurement of the molecule of interest that only carries a donor fluorophore. The donor fluorescence in the presence of the acceptor, F_{DA}, is measured with the molecule of interest that contains both donor and acceptor fluorophore (Figure 21.74). The FRET efficiency E is then calculated as

$$E = 1 - \frac{F_{DA}}{F_D} \qquad \text{eq. 21.141}$$

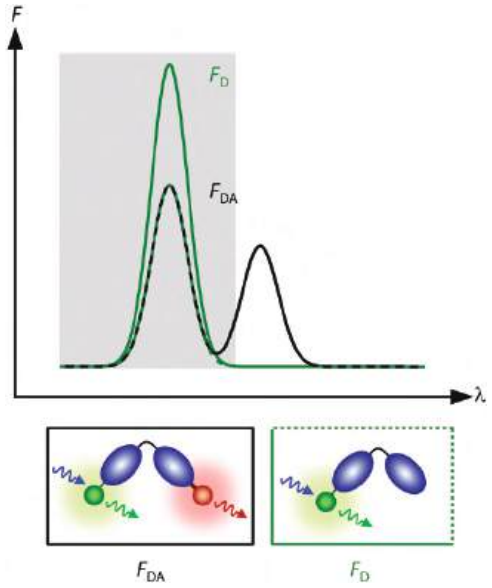


FIGURE 21.74: FRET from donor quenching. The fluorescence emission spectra of the molecule carrying only a donor (F_D, green) and of the molecule carrying donor and acceptor (F_{DA}, black) after excitation of the donor are measured. To extract the donor contribution from the spectrum of the donor/acceptor-labeled molecule, the spectrum is described as the weighted sum of the donor spectrum (green dotted line) and the acceptor spectrum (not shown). The FRET efficiency can be calculated from the integrated intensities F_{DA} (area under the dotted green curve) and F_D (area under the green curve; eq. 21.141).

To determine the FRET efficiency from sensitized acceptor emission, the fluorescence intensity of the acceptor in the absence of donor, F_A, is measured from molecules carrying the acceptor fluorophore only (Figure 21.75).

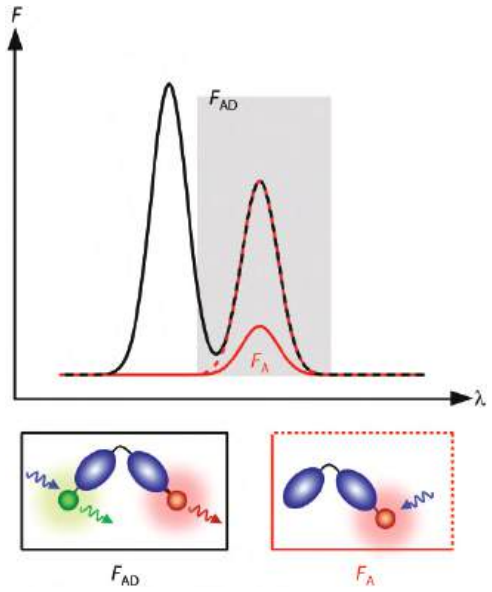


FIGURE 21.75: FRET from acceptor sensitization. To determine the FRET efficiency from the intensity of sensitized acceptor emission, the fluorescence spectra of the donor/acceptor-labeled molecule (F_{DA}, black) and of the molecule carrying only an acceptor (F_A, red) are measured. The acceptor contribution is extracted from the spectrum of the donor/acceptor-labeled molecule by describing it as the weighted sum of the donor spectrum (not shown) and the acceptor spectrum (red dotted line). The FRET efficiency can then be calculated from the integrated intensities F_{AD} (area under the dotted red curve) and F_A (area under the red curve) according to eq. 21.142.

The fluorescence intensity of the acceptor in the presence of donor, F_{AD}, (not to be confused with F_{DA}, the fluorescence intensity of the donor in the presence of the acceptor, in eq. 21.141!) is measured from molecules carrying both fluorophores:

$$E = \frac{1}{f_D} \cdot \frac{\varepsilon_A}{\varepsilon_D} \left(\frac{F_{AD}}{F_A} - 1 \right) \qquad \text{eq. 21.142}$$

Here, f_D is the fraction of molecules carrying a donor moiety, ε_A and ε_D are the extinction coefficients of the acceptor and the donor fluorophores at the excitation wavelength (of the donor), and F_A and F_{AD} are the fluorescence intensities of the acceptor in the absence and presence of the donor fluorophore. The emission of the acceptor fluorophore needs to be corrected for contributions from the fluorescence of the donor, which can be determined from the reference spectrum of molecules carrying a donor only.

Regardless of the method of calculation, fluorescence intensities of ensembles of molecules measured under steady-state illumination are averages over all molecules present. Naturally, FRET efficiencies calculated from average fluorescence intensities represent an average FRET efficiency. Distances obtained from these values by eq. 21.134 may not necessarily represent a true intramolecular distance but an average of all distances present in the sample over the time interval used for the measurement.

21.5.10.3 Applications

Intramolecular FRET can be used as a probe for conformational changes of biomolecules. FRET has been employed to study protein folding, ligand-induced conformational changes, and to follow secondary and tertiary structure formation of nucleic acids (Figure 21.76). Intermolecular FRET serves as a probe for binding (Figure 21.76, Box 21.17).

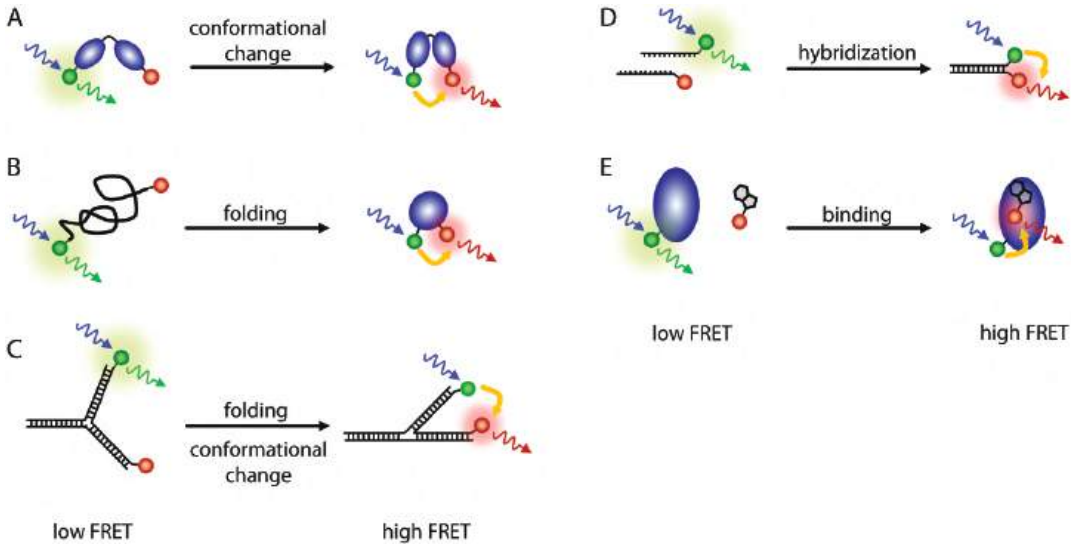


FIGURE 21.76: Applications of FRET. FRET can be used to monitor conformational changes and folding. Intramolecular FRET follows changes in FRET between donor (green) and acceptor (red) dyes that are attached to the same molecule (A-C). Intermolecular FRET follows changes in FRET between dyes that are attached to two binding partners (D, E). A: Protein conformational changes, B: Protein folding, C: Nucleic acid folding and conformational changes, D: DNA hybridization, E: Ligand binding.

BOX 21.17: USING ENERGY TRANSFER FROM TRYPTOPHAN TO MANT AS A PROBE FOR BINDING OF MANT NUCLEOTIDES.

The emission spectrum of Trp (λ_{max} = 320–350 nm) overlaps with the absorption spectrum of mant nucleotides (λ_{max} = 360 nm). As a consequence, Trp can serve as a donor for energy transfer to the mant group. FRET from Trp to the mant group provides a highly selective probe to monitor nucleotide binding. In such an experiment, Trp fluorescence is excited at 280–295 nm, where mant does not show absorption. Mant fluorescence emission is monitored at 440 nm. Under these conditions, the observed fluorescence signal is caused only by protein-nucleotide complexes, with no contribution from free mant nucleotide.

FRET can also be used for fluorescence sensing. FRET is a ratiometric method because both donor and acceptor emission are measured, and is thus concentration-independent. Calmodulin, a Ca^{2+}-binding protein, undergoes a conformational change upon Ca^{2+} binding. By coupling the fluorescent proteins CFP and YFP to the termini of Calmodulin, this conformational change can be detected by a change in the FRET efficiency between the fluorescent proteins. This fluorescence sensor offers the advantage that it can be genetically encoded and produced directly in the cell. Numerous FRET sensors have been developed to sense ions, metabolites, and signaling molecules.

In addition to using FRET as a qualitative probe, FRET can be employed for the determination of distances according to eq. 21.134. These distances can then be used for mapping of conformational states of proteins, or to determine the architecture of large protein complexes (Box 21.18).

BOX 21.18: DETERMINING THE STRUCTURAL ORGANIZATION OF THE RNA POLYMERASE-PROMOTOR OPEN COMPLEX BY FRET.

The architecture of the bacterial RNA polymerase holoenzyme (RNA polymerase and initiation factor σ^{70}) and of the promotor open complex (RNA polymerase/σ^{70}/DNA template) was defined in FRET experiments by determination of individual donor-acceptor distances and triangulation and distance-constrained docking of the individual structures. RNA polymerase was labeled with the donor fluorescein in four different positions, using intein-mediated C-terminal labeling of the polymerase subunits: N-terminal fragments of the subunit up to the amino acid where the label was to be introduced were produced as C-terminal intein fusions, cleaved to generate a C-terminal thioester, and coupled to cysteinylamido-acetamido-fluorescein. The C-terminal part of the subunit was provided as a separate polypeptide during reconstitution of RNA polymerase. The positions for labeling were distant from each other, chosen to ensure that the position of an acceptor dye on σ^{70} would be well defined by the triangulation procedure. The reconstituted labeled polymerase variants were tested for interaction with σ^{70}, and their transcription activity was confirmed. The acceptor fluorophore tetramethylrhodamine (TMR) was introduced at 18 different sites in σ^{70} by Cys-specific labeling with a TMR-maleimide derivative. All labeled σ^{70} variants showed interaction with RNA polymerase and were used in FRET experiments of the holoenzyme. 15 of 18 labeled σ^{70} variants still fully supported transcription, and were used for FRET experiments of the promotor open complex. Labeled DNA was generated by synthesis of Cy5-labeled oligonucleotides that were used as primers in PCR reactions to generate a DNA template with Cy5 in the +15, +20, and +25 position from the transcription

(*Continued*)

start site. Holoenzyme and promotor open complexes were assembled and purified by native polyacrylamide gel electrophoresis. 66 and 105 pairwise FRET efficiencies between fluorescein and TMR and between TMR and Cy5 were determined in steady-state fluorescence experiments of holoenzyme or promotor open complex in gel slices. Steady-state and time-resolved fluorescence anisotropy measurements were performed to examine the local mobility of the dyes and to confirm that the dyes reorient on the time scale of the excited state and that the $\kappa^2 = ⅔$ approximation is valid. From the measured FRET efficiencies, mean inter-dye distances were calculated, and the distances were used as restraints for docking of crystal structures and homology models of individual subunits and fragments thereof to generate structural models for the holoenzyme (RNA polymerase/σ^{70}) and the promotor open complex (RNA polymerase/σ^{70}/DNA template).

The structural models that were obtained from FRET-constrained docking were well defined, with low root mean square distances between individual models. When 10% of the restraints were omitted in the docking procedure, the structures that were obtained were very similar, indicating the robustness of the procedure. The structural models rationalize biochemical and genetic data on the interaction between σ^{70} and the polymerase core, and between σ^{70} and the DNA template in the promotor open complex. Furthermore, they suggest a mechanism for the reduced σ^{70} affinity for the polymerase core after synthesis of a 9-mer to 11-mer RNA. The R3.2 domain of σ^{70} occupies the RNA exit channel of the polymerase. Possibly, when a 9–11 nucleotide RNA has been synthesized it competes with this domain for binding to the exit channel, displaces the R3.2 domain and thereby causes the reduced affinity. Mekler *et al.* (2002) Cell 108: 599–614.

FRET is also used in imaging applications (Section 25.7.6). While fluorescence microscopy is typically limited to the localization of biomolecules, FRET provides additional information. Intramolecular FRET serves as a probe for the conformational state of the molecule of interest. Intermolecular FRET, on the other hand, is an indication for colocalization and/or interaction of the labeled molecules. FRET microscopy on the single-molecule level (Section 25.7.6) provides information on the distribution of conformational states in equilibrium, and on the rates of their interconversion.

21.5.10.4 Potential pitfalls

FRET is a robust qualitative probe for changes in distance between donor and acceptor fluorophores due to conformational rearrangements or binding. It is advisable to always measure changes in both donor and acceptor fluorescence. Anti-correlated changes of donor and acceptor fluorescence are a clear indication of a change in FRET, whereas changes that are not anti-correlated point to effects other than FRET on donor and acceptor fluorescence.

Preparation of donor/acceptor-labeled molecules is the central challenge in intramolecular FRET studies. Ideally, donor and acceptor fluorophores should be introduced into the same molecule site-specifically. Although a variety of different labeling techniques have been established (Section 21.5.4), the combination of orthogonal methods for the introduction of donor and acceptor fluorophores is not straightforward, and the optimization of labeling reactions can be tedious. A common approach for donor/acceptor-labeling of proteins is statistical labeling of a variant with two solvent-exposed cysteines with a mixture of donor- and acceptor-maleimides. Such a labeling reaction yields a mixture of donor/donor-, donor/acceptor-, acceptor/donor-, and acceptor/acceptor-labeled molecules. Incomplete labeling further increases the number of different species (D/–, –/D, A/–, –/A). Only the donor/acceptor-labeled species provide FRET data. All other species are unwanted, but contribute to the measured intensities and affect the experimentally determined FRET efficiency. The species carrying donor and acceptor can in principle be isolated by chromatographic purification, although the separation may be challenging as the chromatographic behavior of all species becomes more and more similar the larger the protein under investigation.

The analysis of FRET experiments requires careful control experiments to determine the effect of the local environment on donor quantum yield and on donor and acceptor spectral properties. To convert FRET efficiencies into inter-dye distances, the Förster distance for the particular donor-acceptor pair has to be determined. The position of dye attachment, as well as the presence of binding partners, can affect the photophysical properties of the dyes and their mobility. Restriction of the local mobility of the dye influences the orientation factor κ^2. Changes in these parameters affect the Förster distance. The local mobility of the dye can be determined from fluorescence anisotropy decays (Section 21.5.9.2) to verify that the $\kappa^2 = ⅔$ approximation holds. In FRET experiments aiming at distance determination, the Förster distance has to be determined for each labeling configuration, and under the exact conditions (solvent, temperature, presence of interaction partners) of the FRET experiment. Even careful controls cannot overcome the inherent limitation of ensemble FRET efficiencies, the averaging over all molecules present in the sample.

21.5.10.5 FRET efficiencies from lifetimes

A major drawback of ensemble FRET experiments lies with the averaging over the observed ensemble. For heterogeneous samples, distances calculated from these FRET efficiencies then represent a population-weighted average, and do not reflect a true intramolecular distance. This limitation can be overcome by measuring the fluorescence lifetime of the donor in the absence and presence of the acceptor. FRET provides an additional pathway for depopulation of the excited state of the donor, and therefore reduces the fluorescence lifetime of the donor. Therefore, FRET efficiencies can also be calculated from fluorescence lifetimes of the donor in the absence (τ_D) and presence (τ_{DA}) of the acceptor:

$$E = 1 - \frac{\tau_{DA}}{\tau_D} \qquad \text{eq. 21.143}$$

In eq. 21.143, the lifetime of the fluorophore in the absence and presence of the acceptor is a measure of the steady-state fluorescence intensity (eq. 21.141). Eq. 21.143 can be applied if the donor shows a single fluorescence lifetime. If more than one lifetime is present, the (intensity-weighted) average lifetime has to be used to calculate the FRET efficiency.

Time-resolved fluorescence spectroscopy offers an additional advantage over the calculation of FRET efficiencies from intensities: the duration of the experiment, *i.e.* from the excitation pulse to the first emitted photon, is very short, on the order of nanoseconds. On this time scale, molecules are essentially static, and different conformers do not interconvert. As a consequence, the fluorescence decay contains contributions from all species present, but conserves the information on different donor-acceptor distances in different molecules. The distribution of these distances can be extracted from the measured decays. First, the donor fluorescence decay in the absence of the acceptor is measured. A multi-exponential decay of the donor fluorescence can be described by eq. 21.144.

$$F_D(t) = \sum_i \alpha_i \cdot \exp\left(-\frac{t}{\tau_i}\right) \qquad \text{eq. 21.144}$$

In a second experiment, the fluorescence decay of the donor in the presence of the acceptor is measured. By equating eq. 21.134 and eq. 21.143

$$E = \frac{R_0^6}{R_0^6 + r^6} = 1 - \frac{\tau_{DA}}{\tau_D} \qquad \text{eq. 21.145}$$

and solving for τ_{DA}, we obtain

$$\tau_{DA} = \tau_D \cdot \frac{r^6}{R_0^6 + r^6} \qquad \text{eq. 21.146}$$

or

$$\tau_{DA} = \tau_D \cdot \frac{r^6}{R_0^6 + r^6} = \frac{\tau_D}{\left(1 + \frac{R_0^6}{r^6}\right)} \qquad \text{eq. 21.147}$$

From eq. 21.147 we see that the lifetime τ is shortened by $(1+R_0{}^6/r^6)$ because of energy transfer: we can now use this expression for each τ_i in eq. 21.144 to describe the donor fluorescence decay in the presence of the acceptor, $F_{DA}(t)$ as

$$F_{DA}(t) = \sum_i \alpha_i \cdot \exp\left(-\frac{t}{\tau_i} \cdot \left(1 + \frac{R_0^6}{r^6}\right)\right) \qquad \text{eq. 21.148}$$

Eq. 21.148 is valid if donor and acceptor are separated by a discrete distance r. If we now assume that the molecules are somewhat flexible, allowing the donor-acceptor distance to fluctuate around a mean value, we can describe the distance distribution $P(r)$ as a three-dimensional Gaussian distribution according to

$$P(r) = 4\pi r^2 c \cdot e^{-a(r-b)^2} \qquad \text{eq. 21.149}$$

Here, the parameter a corresponds to the width of the distance distribution, b marks its center, and c is an amplitude factor (Figure 21.77). We can then substitute the discrete distance r in eq. 21.148 with the *distance distribution* $P(r)$, and have to integrate over all distances r of the distribution, leading to eq. 21.150:

$$F_{DA}(t) = \int_0^\infty P(r) \sum_i \alpha_i \cdot \exp\left(-\frac{t}{\tau_i} \cdot \left(1 + \frac{R_0^6}{r^6}\right)\right) dr \qquad \text{eq. 21.150}$$

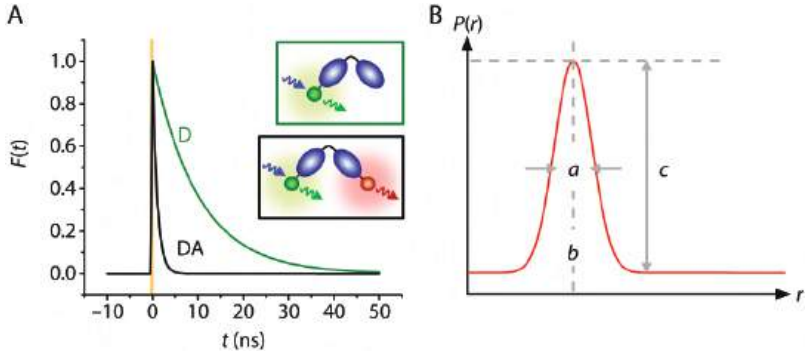


FIGURE 21.77: FRET from fluorescence lifetime measurements. A: The donor fluorescence lifetime τ_D (green) is shortened to τ_{DA} when FRET occurs, and the donor fluorescence decays faster in the presence of the acceptor (black). B: The donor-acceptor distance distribution over all molecules in the solution can be described by a three-dimensional Gaussian distribution (red, eq. 21.149). The parameters *a*, *b*, and *c* describe the width, the center, and the amplitude of the distribution.

Eq. 21.150 holds for a single distance distribution only. For heterogeneous samples with different conformational states and different distance distributions $P_n(r)$, we can generalize eq. 21.150 by summing over n distance distributions with a fractional population f_n:

$$F_{DA}(t) = \sum_n f_n \int P(r) \sum_i \alpha_i \exp\left(-\frac{t}{\tau_i} \cdot \left(1 + \frac{R_0^6}{r^6}\right)\right) dr \qquad \text{eq. 21.151}$$

Thus, by measuring donor fluorescence decays in the absence and presence of acceptor, we can determine the number and fractional populations of the species present (Figure 21.78). The position and width of the Gaussian distribution describe the variations in the distance within one species. From the fractional population of two species in equilibrium, the equilibrium constant can be calculated (Box 21.19). Using time-resolved FRET, it is therefore possible to determine energetic contributions of individual interactions to the stabilization of a conformer by assessing their effect on the conformational equilibrium (Box 21.19).

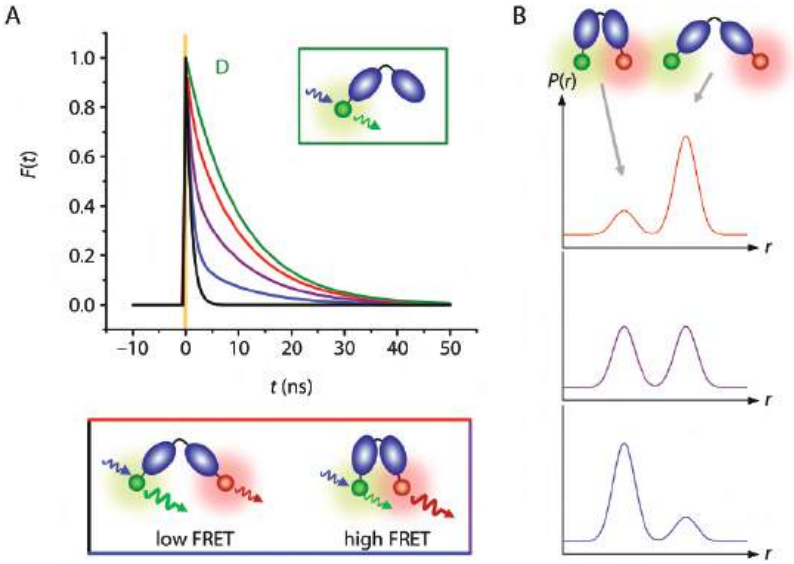


FIGURE 21.78: Conformational states and their population from fluorescence lifetime measurements. Donor fluorescence decays in the presence of the acceptor (red, purple, blue, black) are faster when FRET occurs than in the absence of FRET (green). By description of the measured decays with eq. 21.151, the distance distribution $P(r)$ may be extracted. The two Gaussian distributions represent the donor-acceptor distance in the low FRET species and the high FRET species. From the areas under the curve, the relative population of the two species can be determined. Red 20% high FRET, purple 50% high FRET, blue 80% high FRET, black 100% high FRET.

BOX 21.19: EQUILIBRIUM CONSTANTS FOR CONFORMATIONAL EQUILIBRIA FROM TIME-RESOLVED FRET EXPERIMENTS.

From the donor fluorescence decay in the presence of acceptor, we can calculate distributions of donor-acceptor distances in two (or more) species with different FRET efficiencies. The fractional populations of two conformers in equilibrium, such as a folded and an unfolded state, can thereby be determined (Figure 21.79). From the fractional populations, the equilibrium constant can be calculated, which in turn is linked to ΔG^0 of the reaction (Section 2.1). Energetic contributions of putative interactions that stabilize one conformer over the other can be probed in time-resolved FRET experiments. Donor fluorescence decays for the donor/acceptor-labeled unmodified protein or nucleic acid provide the unperturbed equilibrium constant K_1 and the energetic difference ΔG^0_1 between the two conformers. In a second experiment, the donor decay is measured for a modified version in which a group that is hypothesized to be involved in a stabilizing interaction has been removed. For example, the possible energetic contribution of a hydrogen bond formed by a tyrosine hydroxyl group to the overall protein stability can be probed by replacement of this tyrosine by a phenylalanine. The equilibrium constant K_2 of folding for this variant is determined by time-resolved FRET experiments, and ΔG^0_2 is calculated. If K_2 differs from K_1, the energetic contribution of the tyrosine for stability can be quantified as a $\Delta\Delta G^0$ value, the difference between ΔG^0_1 and ΔG^0_2 (see Section 2.4, Box 2.3).

(*Continued*)

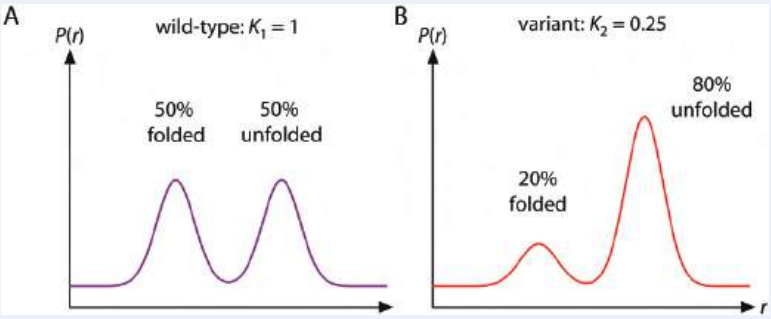


FIGURE 21.79: Energetic contributions of interactions from time-resolved FRET experiments. A. Distance distributions allow quantification of folded (low inter-dye distance) and unfolded forms (high inter-dye distance) of a protein or nucleic acid in equilibrium. The areas under the curve reflect the fractional populations, from which equilibrium constants and ΔG^0 values for folding are calculated. K_1 for folding is 1, the corresponding ΔG^0_1 is 0 kJ mol^{-1}. B: A destabilized variant is only 20% folded under the same conditions. K_2 for folding is 0.25, the corresponding ΔG^0_2 is 3.4 kJ mol^{-1}. The energetic contribution of the group that has been removed in this variant to the overall stability of the folded state is $\Delta\Delta G^0 = \Delta G^0_2 - \Delta G^0_1 = 3.4$ kJ mol^{-1}.

21.5.10.6 FRET efficiencies from single molecules

When donor/acceptor-labeled molecules are studied by single-molecule fluorescence microscopy, FRET efficiencies for single molecules can be determined (see Section 25.7). The FRET efficiency for donor/acceptor-labeled single molecules, either in solution or immobilized on a surface, can be calculated from the measured intensities of the donor and acceptor fluorescence, F_D and F_A, of a single molecule according to eq. 21.138. FRET efficiencies calculated for a large number of single molecules, one after the other (in solution) or in parallel (on surfaces), can be assembled in a FRET histogram that shows how frequently a certain FRET efficiency is encountered. These FRET histograms directly reveal the different populations present, and their fractions, without any assumptions on the shape of the distribution. For surface-immobilized molecules, the donor and acceptor intensities can be measured over seconds to minutes. Changes in the FRET efficiencies in this time window can be followed directly, which enables direct monitoring of conformational changes over time. Single-molecule FRET experiments avoid averaging over ensembles of molecules, and provide important insight into heterogeneous systems with different species of different donor-acceptor distances. They offer the possibility to monitor events that cannot be synchronized. Single-molecule FRET and its applications are discussed in more detail in Section 25.7.

QUESTIONS

21.1 A solution of ATP has a concentration of 100 µM. What is the absorption and transmission (d = 1 cm, ε(ATP) = 15400 M^{-1} cm^{-1})?

21.2 Derive a general formula for calculation of the transmission T in percent from the absorption A. Which absorption results in 50% transmission?

21.3 What is the wavelength of the radio waves used by the Tijuana-based alternative music radio station 91.X, broadcasting to San Diego at a frequency ν = 91.10 MHz? What energy does a single photon carry? By how much would a mole of photons heat a cup of tea (V = 100 mL, C_p = 75 J mol^{-1} K^{-1}), if absorbed completely?

21.4 Proteins typically have an absorption ratio of $A_{280}/A_{260} = 2$, whereas the ratio for nucleic acids and nucleotides is $A_{280}/A_{260} = 0.5$. A purified ATP-binding protein with a molar extinction coefficient $\varepsilon_{280} = 12000\ M^{-1}\ cm^{-1}$, has an absorption of $A_{280} = 0.5$, $A_{260} = 0.35$. What is the fraction of ATP (or ADP) that has been co-purified (ε_{260} (ATP) = 15400 M^{-1} cm^{-1}, $d = 1$ cm)? Calculate the A_{280}/A_{260} ratio for a nucleotide contamination of 1%.

21.5 Fluorescein absorbs light of $\lambda = 495$ nm. What is the energy difference ΔE of ground state S_0 and excited state S_1 per molecule and per mol? The extinction coefficient is $\varepsilon = 95000\ M^{-1}\ cm^{-1}$. How much energy is necessary to excite all molecules in one mL of a fluorescein solution with $A = 1.2$ ($d = 1$ cm)? What fraction of the incident light reaches the detector?

21.6 NADH has an extinction coefficient ε_{340} of 6300 M^{-1} cm^{-1}; NAD^+ does not absorb at this wavelength. At 260 nm, NAD^+ absorbs more strongly, with ε_{260} (NAD^+) = 18000 M^{-1} cm^{-1}, and ε_{260} (NADH) = 15000 M^{-1} cm^{-1}. You find a very old stock of NADH in the freezer, and prepare a solution that has an absorption of $A_{260} = 0.31$, and $A_{340} = 0.11$ ($d = 0.2$ cm). What is the fraction of NAD^+?

21.7 A solution contains glucose-6-phosphate and glucose-1-phosphate. When an excess of $NADP^+$ and glucose-6-phosphate dehydrogenase is added, a final absorption at 340 nm of $A_{340} = 0.60$ is measured. Subsequent addition of phosphoglucomutase leads to an increase in absorption to $A_{340} = 0.72$. What are the glucose-6-phosphate and glucose-1-phosphate concentrations in the solution? ε_{340}(NADPH) = 6300 M^{-1} cm^{-1}, $d = 1$ cm.

21.8 Addition of Ellman's reagent to a 15 µM protein solution leads to an increase in A_{412} by 0.22 units. In denaturing buffer, the increase is 0.43 units. How many surface-exposed and buried cysteines does the protein contain? Does the protein contain disulfide bridges?
$\varepsilon(TNB^-) = 14000\ M^{-1}\ cm^{-1}$.

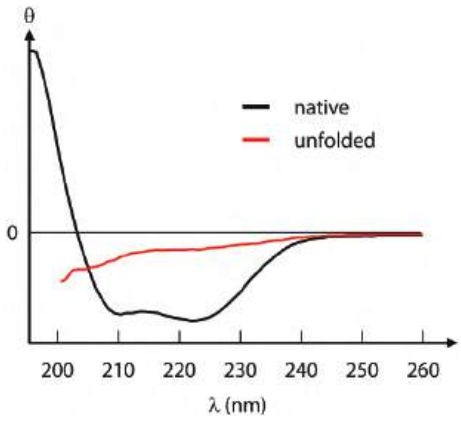
21.9 An excess of Ellman's reagent is added to a 10 µM solution of an enzyme. Under native conditions, the absorption increases by 0.15, under denaturing conditions it increases by 0.29. Calculate the number of accessible cysteines and explain the result. Repeating the experiment in the presence of the substrate, the absorption increases by 0.04 under native, by 0.30 under denaturing conditions. What is the reason for this behavior? What is the degree of binding? $\varepsilon(TNB^-) = 14000\ M^{-1}\ cm^{-1}$.

21.10 How can you determine the concentration of a structured RNA molecule of 154 nucleotides, consisting of 54 A, 35 G, 29 C, and 36 U, most accurately by absorption? How can you determine the extinction coefficient of the folded RNA molecule? Why is the procedure not applicable to DNA?

21.11 A CD measurement of the ellipticity at 222 nm gives $\theta_A = 1.246$ mdeg for a 0.9 µM solution of protein A, a 23.4 kDa protein of 210 amino acids, and $\theta_B = 2.345$ mdeg for a 1.15 µM solution of protein B, a 43.2 kDa protein of 393 amino acids. Which protein has a higher α-helical content? $d = 0.1$ cm.

21.12 A protein folding intermediate shows a high far-UV CD signal, no near-UV CD, and tryptophan fluorescence at 330 nm wavelength. What are the characteristics of the intermediate?

21.13 CD spectra for a protein in its folded and unfolded form are given.



Which wavelength is the most suitable for following unfolding of the protein by CD spectroscopy? Why is the wavelength with maximal signal difference not preferable?

21.14 The heat-induced unfolding of a protein is measured using far-UV CD and tryptophan fluorescence as spectroscopic probes. The melting temperature determined from the transition observed by CD is 5 degrees higher than the melting temperature determined by fluorescence. What is the reason?

21.15 Calculate the stiffness k for CO-, NH-, CH-, and CD- bonds from the harmonic oscillator model. The wavenumbers are 1720 cm^{-1} (CO), 3400 cm^{-1} (NH), 2900 cm^{-1} (CH), and 2200 cm^{-1} (CD). Calculate the average change in bond lengths Δx during the stretch vibrations. Is the isotope effect for CH and CD only caused by the difference in (reduced) mass?

21.16 The wavenumber for the C–F stretch vibration is 1200 cm^{-1}, the wavenumbers for the C–Cl, C–Br, and C–I stretch vibrations are 700 cm^{-1} (C–Cl), 550 cm^{-1} (C–Br), and 500 cm^{-1} (C–I). Can this effect be explained by the different masses of the halogens?

21.17 Why is it necessary to calculate difference spectra of free and bound forms of a fluorescent ligand to select the emission wavelength at which the signal change for a fluorescence titration is maximal? Under what conditions is the observation at the maximum of the emission spectrum optimal?

21.18 A new student in the laboratory measures fluorescence of a concentrated protein solution, and obtains a relative fluorescence of 0.68. A more experienced student dilutes the protein solution 10-fold before the measurement (under otherwise identical conditions) and measures a relative fluorescence of 0.3. The protein has an extinction coefficient of 25000 M^{-1} cm^{-1} at the excitation wavelength, and of 12500 M^{-1} cm^{-1} at the emission wavelength. What is the concentration of the concentrated protein solution? (d= 1 cm).

21.19 Calculate the equations for conversion of anisotropy r into polarization p and *vice versa*.

21.20 A fluorescent ligand in solution has an anisotropy of 0.05. The anisotropy increases to 0.25 upon binding of the ligand to a protein. Calculate the anisotropy and polarization of a solution with 40% saturation. What is the deviation (in %), if a linear dependence of polarization on concentration of bound and free forms is assumed?

21.21 A fluorophore has a quantum yield of 0.2 and a fluorescence lifetime of τ = 12 ns. Calculate the rate constants k_f and k_{nr} for radiative and non-radiative transitions from S_1 to S_0 and the intrinsic lifetime τ_0. Upon addition of 0.25 M acrylamide, the lifetime is reduced to 10 ns. What is the Stern-Volmer constant K_{SV} for this quencher? Calculate the K_d value of the fluorophore complex with a quencher that would achieve the same quenching effect by static quenching. The concentration of the fluorophore is 1 µM.

21.22 A protein is very difficult to purify and tends to aggregate and precipitate when it is titrated to a fluorescent ligand. To be able to determine the K_d value of the protein-ligand complex, the titration is performed in reverse, by supplying the protein at a low concentration, and titrating with increasing concentrations of the fluorescent ligand. Derive the equation for the measured fluorescence F as a function of total ligand and protein concentration and the K_d. Why is the binding curve obtained more difficult to analyze than a direct titration of fluorescent ligand with the protein?

21.23 A double-cysteine variant of a protein ($\varepsilon_{protein,\ 280\ nm}$ = 14800 M^{-1} cm^{-1}) has been labeled with a mixture of donor and acceptor maleimides for FRET experiments ($\varepsilon_{donor,\ 495\ nm}$ = 72500 M^{-1} cm^{-1}, $\varepsilon_{acceptor,\ 540\ nm}$ = 95000 M^{-1} cm^{-1}). Both donor and acceptor fluorophores show absorption at 280 nm ($\varepsilon_{donor,\ 280\ nm}$ = 9300 M^{-1} cm^{-1}, $\varepsilon_{acceptor,\ 280\ nm}$ = 26500 M^{-1} cm^{-1}), and the acceptor absorbs at the excitation wavelength of the donor ($\varepsilon_{acceptor,\ 495\ nm}$ = 18600 M^{-1} cm^{-1}). Determine the labeling efficiencies for donor and acceptor from the absorption values at 280 nm, 495 nm, and 540 nm (d = 1 cm), A_{280} = 0.25, A_{495} = 0.31, A_{540} = 0.49. What is the fraction of unlabeled molecules and of protein molecules that carry one donor, two donors, one acceptor, two acceptors, one donor and one acceptor, or no label at all?

21.24 A colleague determines the K_d value of a DNA-protein complex by titration of the protein with a 10-base-pair DNA, and follows binding by measuring the fluorescence of intrinsic tryptophan residues. After an initial increase in tryptophan fluorescence, the fluorescence decreases linearly until it reaches background intensity. What is the reason for this behavior? How can the measured fluorescence values be corrected to obtain a binding curve?

21.25 A globular protein is labeled with fluorophore A that has a lifetime τ_A = 1 ns, or fluorophore B with a lifetime τ_B = 10 ns. Which of the labeled proteins has the higher steady-state anisotropy and why?

21.26 Calculate the steady-state anisotropies for fluorophores with lifetimes τ = 0.1 ns, 0.5 ns, 1 ns, 5 ns, 10 ns, 50 ns, and 100 ns that are rigidly attached to a molecule with τ_c = 1 ns. Binding of the molecule to a protein causes an increase in rotational correlation time to τ_c = 10 ns. Which fluorophore is best suited to follow binding, which fluorophores do not report on binding?

21.27 Two proteins of 25 kDa (A) and 45 kDa (B) form a heterodimer. One protein has no tryptophan, the other contains a single tryptophan. What are the prerequisites for measuring binding by a change in tryptophan anisotropy?

21.28 Fluorescein has a lifetime of τ_D = 5.2 ns when attached to a DNA. When tetramethylrhodamine is attached at the other end of the DNA, the lifetime decreases to τ_{DA} = 3.6 ns. Upon binding of a protein to the DNA, the lifetime of the donor decreases to 4 ns, the lifetime of the donor in the presence of acceptor decreases to 1.2 ns. What is the energy transfer efficiency, the rate constant k_t for energy transfer, and the donor-acceptor distance for the free DNA and for the protein-DNA complex? R_0 = 5.4 nm.

21.29 A Cy3-Cy5-labeled protein occurs in three different conformational states with inter-dye distances of 1 nm (50%), 2 nm (10%), and 5 nm (40%). The percentages are the fraction of each conformer in equilibrium. The Förster distance is R_0 = 5.2 nm. Calculate the FRET efficiency for each conformer, the mean FRET efficiency, and the average distance over all conformers. What is the mean distance that would be determined from an ensemble FRET experiment? What additional information would be obtained by time-resolved FRET?

21.30 A fluorescein-TMR-labeled enzyme is analyzed with time-resolved FRET. The donor lifetime is 5 ns in the absence and 4 ns in the presence of acceptor. When substrate is added, the donor lifetime is 5 ns in the absence and 1 ns in the presence of the acceptor. The Förster distance is R_0 = 5.2 nm. Calculate E_{FRET} and r for both cases.

REFERENCES

Allin, C., Ahmadian, M. R., Wittinghofer, A. and Gerwert, K. (2001) Monitoring the GAP catalyzed H-Ras GTPase reaction at atomic resolution in real time. *Proc Natl Acad Sci U S A* **98**(14): 7754–7759.
· application of step-scan FT-IR spectroscopy to the mechanism of GTP hydrolysis by Ras in presence of the Ras GTPase-activating protein

Andreou, A. Z. and Klostermeier, D. (2014) eIF4B and eIF4G jointly stimulate eIF4A ATPase and unwinding activities by modulation of the eIF4A conformational cycle. *J Mol Biol* **426**(1): 51–61.
· use of 2-aminopurine to monitor RNA unwinding by a helicase

Andreou, A. Z. and Klostermeier, D. (2014) Fluorescence methods in the investigation of the DEAD-box helicase mechanism. *EXS* **105**: 161–192.
· review article on the application of fluorescence techniques to DEAD-box helicases

Appel, M. J. and Bertozzi, C. R. (2015) Formylglycine, a post-translationally generated residue with unique catalytic capabilities and biotechnology applications. *ACS Chem Biol* **10**(1): 72–84.
· review about formylglycine generating enzyme that can be used to genetically encode formyl groups

Asmari, M., Michalcová, L., Ibrahim, A. E., Glatz, Z., Wätzig, H. and El Deeb, S. (2023) Studying molecular interactions *via* capillary electrophoresis and microscale thermophoresis: A review. *Electrophoresis* **44**(13–14): 1114–1142.
· review summarizing recent applications of microscale thermophoresis to a variety of biomolecular interactions

Bayley, P. M. and Harris, H. E. (1975) Conformational properties of pig-heart cytoplasmic aspartate aminotransferase. Circular-dichroism and absorption-spectroscopic study of dicarboxylate binding. *Eur J Biochem* **56**(2): 455–465.
· study on the induced circular dichroism of pyridoxal phosphate bound to aspartate aminotransferase

Birdsall, B., King, R. W., Wheeler, M. R., Lewis, C. A., Jr., Goode, S. R., Dunlap, R. B. and Roberts, G. C. (1983) Correction for light absorption in fluorescence studies of protein-ligand interactions. *Anal Biochem* **132**(2): 353–361.
· correction of absorption values for inner-filter effects

Brannon, J. H. and Magde, D. (1978) Absolute quantum yield determination by thermal blooming. Fluorescein. *J Phys Chem* **82**(6): 705–709.
· determination of fluorescein quantum yield in 0.1 M NaOH by thermal analysis

Bulheller, B. M. and Hirst, J. D. (2009) DichroCalc-circular and linear dichroism online. *Bioinformatics* **25**(4): 539–540.
· description of the program DichroCalc for calculations of circular and linear dichroism spectra

Cabantous, S., Nguyen, H. B., Pedelacq, J. D., Koraichi, F., Chaudhary, A., …, Waldo, G. S. (2013) A new protein-protein interaction sensor based on tripartite split-GFP association. *Sci Rep* **3**: 2854.
· generation of tripartite GFP to study protein-protein interactions

Cabantous, S., Terwilliger, T. C. and Waldo, G. S. (2005) Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotechnol* **23**(1): 102–107.
· generation of self-assembling GFP fragments

Cantor, C. R. and Schimmel, P. R. (1980) *Biophysical Chemistry (Part II) Techniques for the Study of Biological Structure and Function*. New York, Freeman.
· a comprehensive book on biophysical techniques including optical spectroscopy, with a strong focus on physical background and quantitative derivations

Carrico, I. S., Carlson, B. L. and Bertozzi, C. R. (2007) Introducing genetically encoded aldehydes into proteins. *Nat Chem Biol* **3**(6): 321–322.
· original paper that reports the genetically encoded incorporation of formylglycine into proteins

Cepus, V., Scheidig, A. J., Goody, R. S. and Gerwert, K. (1998) Time-resolved FTIR studies of the GTPase reaction of H-ras p21 reveal a key role for the β-phosphate. *Biochemistry* **37**(28): 10263–10271.
· study of the mechanism of GTP hydrolysis by the oncoprotein Ras by Fourier transform infrared spectroscopy using GTP with 18O substitutions at the α-, β-, or γ-phosphate

Chou, P. Y. and Fasman, G. D. (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochemistry* **13**(2): 211–222.
· reports how the CD spectra of proteins depend on the secondary structure

Cohen, A. E. (2016) Optogenetics: Turning the microscope on its head. *Biophys J* **110**(5): 997–1003.
· review on current state of the field of optogenetics

Cordes, T., Vogelsang, J. and Tinnefeld, P. (2009) On the mechanism of Trolox as antiblinking and antibleaching reagent. *J Am Chem Soc* **131**(14): 5018–5019.
· suppression of fluorophore photobleaching and blinking by the vitamin E analog Trolox

Dafforn, T. R. and Rodger, A. (2004) Linear dichroism of biomolecules: Which way is up? *Curr Opin Struct Biol* **14**(5): 541–546.
· review article on application of linear dichroism to biomolecules

Dixon, J. M., Taniguchi, M. and Lindsey, J. S. (2005) PhotochemCAD 2: A refined program with accompanying spectral databases for photochemical calculations. *Photochem Photobiol* **81**(1): 212–213.
· a program for spectral calculations with an accompanying database of absorption and fluorescence spectra

Duhr, S., Arduini, S. and Braun, D. (2004) Thermophoresis of DNA determined by microfluidic fluorescence. *Eur Phys J E Soft Matter* **15**(3): 277–286.
· original paper describing microscale thermophoresis of DNA

Duhr, S. and Braun, D. (2006) Why molecules move along a temperature gradient. *Proc Natl Acad Sci U S A* **103**(52): 19678–19682.
· original paper relating thermophoresis in solution to solvation entropy

Ehresmann, B., Imbault, P. and Weil, J. H. (1973) Spectrophotometric determination of protein concentration in cell extracts containing tRNAs and rRNAs. *Anal Biochem* **54**(2): 454–463.
· photometric determination of protein concentrations in the presence of nucleic acids

Ellman, G. L. (1959) Tissue sulfhydryl groups. *Arch Biochem Biophys* **82**(1): 70–77.
· original paper describing the quantification of cysteines by photometry using DTNB

Fasman, G. D. (1996) *Circular Dichroism and the Conformational Analysis of Biomolecules*. New York, Plenum Press.
· a comprehensive book about theory and applications of circular dichroism of proteins and other biomolecules

Förster, T. (1948) Zwischenmolekulare Energiewanderung und Fluoreszenz. *Annalen der Physik* **437**(1): 55–75.
· original description of the phenomenon of Förster Resonance Energy Transfer

Gauchet, C., Labadie, G. R. and Poulter, C. D. (2006) Regio- and chemoselective covalent immobilization of proteins through unnatural amino acids. *J Am Chem Soc* **128**(29): 9274–9275.
· use of farnesyl transferases to introduce four amino acid sequences with alkyne groups into proteins and modification by click chemistry.

Gautier, A., Juillerat, A., Heinis, C., Correa, I. R., Jr., Kindermann, M., Beaufils, F. and Johnsson, K. (2008) An engineered protein tag for multiprotein labeling in living cells. *Chem Biol* **15**(2): 128–136.
· description of the CLIP tag for fluorescent labeling of proteins *in vivo*

Gerwert, K., Souvignier, G. and Hess, B. (1990) Simultaneous monitoring of light-induced changes in protein side-group protonation, chromophore isomerization, and backbone motion of bacteriorhodopsin by time-resolved Fourier-transform infrared spectroscopy. *Proc Natl Acad Sci U S A* **87**(24): 9774–9778.
· application of step-scan FT-IR spectroscopy to investigate structural changes in bacteriorhodopsin during its photocycle

Gill, S. C. and von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* **182**(2): 319–326.
· photometric method to determine the molar extinction coefficient of folded proteins

Greenfield, N. J. (2006) Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* **1**(6): 2876–2890.
· review article focusing on practical aspects of CD measurements and the interpretation of CD spectra for secondary structure prediction of proteins

Griffin, B. A., Adams, S. R. and Tsien, R. Y. (1998) Specific covalent labeling of recombinant protein molecules inside live cells. *Science* **281**(5374): 269–272.
· original paper describing *in vivo* labeling of proteins using a genetically encoded tetracysteine motif that forms a biarsenic complex with fluorescein

Griffin, B. A., Adams, S. R., Jones, J. and Tsien, R. Y. (2000) Fluorescent labeling of recombinant proteins in living cells with FlAsH. *Methods Enzymol* **327**: 565–578.
· method review of the use of FlAsH tags for fluorescent labeling *in vivo*

Guimaraes, C. P., Witte, M. D., Theile, C. S., Bozkurt, G., Kundrat, L., Blom, A. E. and Ploegh, H. L. (2013) Site-specific C-terminal and internal loop labeling of proteins using sortase-mediated reactions. *Nat Protoc* **8**(9): 1787–1799.
· protocol for C-terminal and internal loop labeling of proteins using sortase

Hochreiter, B., Garcia, A. P. and Schmid, J. A. (2015) Fluorescent proteins as genetically encoded FRET biosensors in life sciences. *Sensors* **15**(10): 26281–26314.
· review on the principle of FRET sensors

Hofmann, C., Ketelaars, M., Matsushita, M., Michel, H., Aartsma, T. J. and Kohler, J. (2003) Single-molecule study of the electronic couplings in a circular array of molecules: Light-harvesting-2 complex from *Rhodospirillum molischianum*. *Phys Rev Lett* **90**(1): 013004.
· determination of the relative orientation of chlorophyll molecules in a bacterial photosynthetic light harvesting complex through their transition dipole moments

Holz, B., Klimasauskas, S., Serva, S. and Weinhold, E. (1998) 2-Aminopurine as a fluorescent probe for DNA base flipping by methyltransferases. *Nucleic Acids Res* **26**(4): 1076–1083.
· investigation of base-flipping by the methyltransferase M.HhaI using 2-AP fluorescence as a probe

Johnson, W. C., Jr. (1990) Protein secondary structure and circular dichroism: A practical guide. *Proteins* **7**(3): 205–214.
· introductory review on practical aspects in the application of circular dichroism to proteins

Jones, A. C. and Neely, R. K. (2015) 2-Aminopurine as a fluorescent probe of DNA conformation and the DNA-enzyme interface. *Q Rev Biophys* **48**(2): 244–279.
· review on applications of 2-aminopurine to study nucleic acid structure and dynamics, with a focus on time-resolved fluorescence

Jung, C. (2000) Insight into protein structure and protein-ligand recognition by Fourier transform infrared spectroscopy. *J Mol Recognit* **13**(6): 325–351.
· review of applications of FT-IR spectroscopy to proteins, with a focus on heme proteins

Keppler, A., Gendreizig, S., Gronemeyer, T., Pick, H., Vogel, H. and Johnsson, K. (2003) A general method for the covalent labeling of fusion proteins with small molecules *in vivo*. *Nat Biotechnol* **21**(1): 86–89.
· description of the SNAP tag for fluorescent labeling of proteins *in vivo*

Kirchhofer, A., Helma, J., Schmidthals, K., Frauer, C., Cui, S., …, Rothbauer, U. (2010) Modulation of protein properties in living cells using nanobodies. *Nat Struct Mol Biol* **17**(1): 133–138.
· generation of nanobodies that bind to green fluorescent protein (GFP), structural basis of the reduction or increase in GFP fluorescence by these nanobodies (GFP minimizer and GFP maximizer), and application in expression and localization studies

Klonoff, D. C. (2012) Overview of fluorescence glucose sensing: A technology with a bright future. *J Diabetes Sci Technol* **6**(6): 1242–1250.
· review about glucose sensing by fluorescence sensors

Kotting, C. and Gerwert, K. (2005) Proteins in action monitored by time-resolved FTIR spectroscopy. *Chemphyschem* **6**(5): 881–888.
· review article on time-resolved FT-IR spectroscopy

Lakowicz, J. R. (1999) *Principles of Fluorescence Spectroscopy*. New York, Kluwer Academic/Plenum Publishers.
· a compendium of fluorescence spectroscopy

Lee, M. H., Kim, J. S. and Sessler, J. L. (2015) Small molecule-based ratiometric fluorescence probes for cations, anions, and biomolecules. *Chem Soc Rev* **44**(13): 4185–4191.
· review article on ratiometric fluorescence sensors

Lentz, B. R. (1993) Use of fluorescent probes to monitor molecular order and motions within liposome bilayers. *Chem Phys Lipids* **64**(1–3): 99–116.
· review article on the use of fluorescent probes to study membrane dynamics and order

Liu, C. C. and Schultz, P. G. (2010) Adding new chemistries to the genetic code. *Annu Rev Biochem* **79**: 413–444.
· review article about incorporation of non-natural amino acids into proteins in *E. coli*, yeast and mammalian cells.

Los, G. V., Encell, L. P., McDougall, M. G., Hartzell, D. D., Karassina, N., …, Wood, K. V. (2008) HaloTag: A novel protein labeling technology for cell imaging and protein analysis. *ACS Chem Biol* **3**(6): 373–382.
· description of the Halo tag for fluorescent labeling

Magde, D., Wong, R. and Seybold, P. G. (2002) Fluorescence quantum yields and their relation to lifetimes of rhodamine 6G and fluorescein in nine solvents: Improved absolute standards for quantum yields. *Photochem Photobiol* **75**(4): 327–334.
· calorimetric study to determine rates of radiative and non-radiative decay for fluorescein and rhodamine 6G in different solvents with high precision

Magliery, T. J., Wilson, C. G., Pan, W., Mishler, D., Ghosh, I., Hamilton, A. D. and Regan, L. (2005) Detecting protein-protein interactions with a green fluorescent protein fragment reassembly trap: Scope and mechanism. *J Am Chem Soc* **127**(1): 146–157.
· use of split versions of GFP to detect protein-protein interactions

Mekler, V., Kortkhonjia, E., Mukhopadhyay, J., Knight, J., Revyakin, A., …, Ebright, R. H. (2002) Structural organization of bacterial RNA polymerase holoenzyme and the RNA polymerase-promoter open complex. *Cell* **108**(5): 599–614.
· structural modeling of the RNA polymerase holoenzyme and the promotor open complex using 66 and 105 FRET-derived distance restraints

Moore, M. J. and Query, C. C. (2000) Joining of RNAs by splinted ligation. *Methods Enzymol* **317**: 109–123.
· description of RNA ligation techniques with a focus on practical aspects

Muir, T. W., Sondhi, D. and Cole, P. A. (1998) Expressed protein ligation: A general method for protein engineering. *Proc Natl Acad Sci U S A* **95**(12): 6705–6710.
· original paper reporting the use of expressed protein ligation for protein engineering

Muralidharan, V. and Muir, T. W. (2006) Protein ligation: An enabling technology for the biophysical analysis of proteins. *Nat Methods* **3**(6): 429–438.
· review article about applications of expressed protein ligation in biophysical analyses

Nagel, G., Ollig, D., Fuhrmann, M., Kateriya, S., Musti, A. M., Bamberg, E. and Hegemann, P. (2002) Channelrhodopsin1: A light-gated proton channel in green algae. *Science* **296**(5577): 2395–2398.
· discovery of channelrhodopsins, light-gated ion channels in algae

Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. and Schultz, P. G. (1989) A general method for site-specific incorporation of non-natural amino acids into proteins. *Science* **244**(4901): 182–188.
· incorporation of non-natural amino acids by suppression of stop codons

Ormo, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y. and Remington, S. J. (1996) Crystal structure of the *Aequorea victoria* green fluorescent protein. *Science* **273**(5280): 1392–1395.
· crystal structure of green fluorescent protein

Paige, J. S., Wu, K. Y. and Jaffrey, S. R. (2011) RNA mimics of green fluorescent protein. *Science* **333**(6042): 642–646.
· development of the green fluorescent RNA aptamer "spinach"

Plach, M. G., Grasser, K. and Schubert, T. (2017) MicroScale thermophoresis as a tool to study protein-peptide interactions in the context of large eukaryotic protein complexes. *Bio Protoc* **7**(23): e2632.
· guide for planning, execution, interpretation and trouble-shooting of microscale thermophoresis experiments

Popp, M. W., Antos, J. M., Grotenbreg, G. M., Spooner, E. and Ploegh, H. L. (2007) Sortagging: A versatile method for protein labeling. *Nat Chem Biol* **3**(11): 707–708.

· description of the use of sortase to attach fluorescent labels to proteins *in vitro*, in cell lysates and on the cell surface

Rabuka, D., Rush, J. S., deHart, G. W., Wu, P. and Bertozzi, C. R. (2012) Site-specific chemical protein conjugation using genetically encoded aldehyde tags. *Nat Protoc* **7**(6): 1052–1067.
· protocol for incorporation of aldehyde groups into proteins using the formylglycine generating enzyme

Rathod, R., Kang, Z., Hartson, S. D., Kumauchi, M., Xie, A. and Hoff, W. D. (2012) Side-chain specific isotopic labeling of proteins for infrared structural biology: The case of ring-D4-tyrosine isotope labeling of photoactive yellow protein. *Protein Expr Purif* **85**(1): 125–132.
· incorporation of isotope-labeled Tyr into photoactive yellow protein for isotope-edited FTIR spectroscopy

Ritzefeld, M. (2014) Sortagging: A robust and efficient chemoenzymatic ligation strategy. *Chem Eur J* **20**(28): 8516–8529.
· review on applications of sortase-mediated tagging

Smith, G. J., Sosnick, T. R., Scherer, N. F. and Pan, T. (2005) Efficient fluorescence labeling of a large RNA through oligonucleotide hybridization. *RNA* **11**(2): 234–239.
· introduction of donor and acceptor fluorophores into RNase P RNA by hybridization of fluorescently labeled oligonucleotides to unpaired bases in loop regions

Strack, R. L. and Jaffrey, S. R. (2013) New approaches for sensing metabolites and proteins in live cells using RNA. *Curr Opin Chem Biol* **17**(4): 651–655.
· spinach derivatives for metabolite sensing in living cells

Stryer, L. and Haugland, R. P. (1967) Energy transfer: A spectroscopic ruler. *Proc Natl Acad Sci U S A* **58**(2): 719–726.
· experimental demonstration of fluorescence resonance energy transfer as a spectroscopic ruler using poly-lysine peptides of different lengths

Sustarsic, M. and Kapanidis, A. N. (2015) Taking the ruler to the jungle: Single-molecule FRET for understanding biomolecular structure and dynamics in live cells. *Curr Opin Struct Biol* **34**: 52–59.
· review article on quantitative *in vivo* applications of FRET

Taki, M., Shiota, M. and Taira, K. (2004) Transglutaminase-mediated N- and C-terminal fluorescein labeling of a protein can support the native activity of the modified protein. *Protein Eng Des Sel* **17**(2): 119–126.
· description of fluorescent labeling by transglutaminase-mediated modification of a glutamine

Theile, C. S., Witte, M. D., Blom, A. E., Kundrat, L., Ploegh, H. L. and Guimaraes, C. P. (2013) Site-specific N-terminal labeling of proteins using sortase-mediated reactions. *Nat Protoc* **8**(9): 1800–1807.
· protocol for N-terminal labeling of proteins using sortase

Thrall, S. H., Reinstein, J., Wohrl, B. M. and Goody, R. S. (1996) Evaluation of human immunodeficiency virus type 1 reverse transcriptase primer tRNA binding by fluorescence spectroscopy: Specificity and comparison to primer/template binding. *Biochemistry* **35**(14): 4609–4618.
· solution of the cubic equation and analysis of displacement titrations

Toseland, C. and Fili, N. (2014) *Fluorescent Methods Applied to Molecular Motors: From Single Molecules to Whole Cells*. Basel, Springer.
· a collection of articles on different fluorescent methods and their application to ATP-dependent enzymes

Vogelsang, J., Kasper, R., Steinhauer, C., Person, B., Heilemann, M., Sauer, M. and Tinnefeld, P. (2008) A reducing and oxidizing system minimizes photobleaching and blinking of fluorescent dyes. *Angew Chem Int Ed* **47**(29): 5465–5469.
· suppression of photobleaching and blinking of fluorophores by reducing/oxidizing systems

Walter, N. G., Harris, D. A., Pereira, M. J. and Rueda, D. (2001) In the fluorescent spotlight: Global and local conformational changes of small catalytic RNAs. *Biopolymers* **61**(3): 224–242.
· review on fluorescence probes for large scale and local conformational changes of RNA, including 2-aminopurine and FRET

Wienken, C. J., Baaske, P., Braun, D. and Duhr, S. (2010) Protein-binding assays in biological liquids using microscale thermophoresis. *Nat Comm* **1**: 100.
· describes the application of microscale thermophoresis to interactions of proteins in buffer and human serum

Wilhelmsson, L. M. (2010) Fluorescent nucleic acid base analogues. *Q Rev Biophys* **43**(2): 159–183.
· review article summarizing the structures and properties of fluorescent nucleotide analogs and possible applications

22 Magnetic resonance

Atoms consist of nuclei and electrons that carry positive and negative charges, respectively. Movement of electric charges generates a magnetic field. Electrons and nuclei can therefore have a magnetic moment, originating from a circular electric current, and can be regarded as tiny magnetic dipoles. This magnetic moment is related to the quantum phenomenon called *spin*. The spins of electrons and nuclei are quantized, and are characterized by the *spin quantum number*. In the presence of external magnetic fields, the magnetic moments adopt preferred orientations relative to the field of different energies. In magnetic resonance methods, transitions between these states are induced by electromagnetic radiation. Magnetic resonance methods provide information on the local environment and mobility of the probes and on inter-probe distances.

22.1 NUCLEAR MAGNETIC RESONANCE

Nuclear magnetic resonance (NMR) is based on the inherent magnetic properties of nuclei. Therefore, we will focus on nuclear spins in the following Sections, and will return to electron spins when we discuss electron paramagnetic resonance (EPR, Section 22.2).

22.1.1 Nuclear spins and the Zeeman effect

The positive charge of nuclei is located on the protons, and all nuclei except hydrogen contain more than one proton. The spin quantum numbers I of nuclei are multiples of ½, *i.e.* 0, ½, 1, ³⁄₂, and so on. In the absence of a magnetic field, the orientation of the magnetic moment of the nuclei is random, and the nuclei are energetically equivalent (degenerate states). In the presence of an external magnetic field, however, the magnetic dipole of a nucleus adopts certain orientations relative to the direction of the field. A nucleus with the spin quantum number I can be found in $(2I+1)$ orientations. NMR experiments on biomolecules are often performed with nuclei with $I = ½$, such as ^{1}H, ^{13}C, ^{15}N, or ^{31}P, whose magnetic moment can adopt only two different orientations relative to the magnetic field. In the simplified image of nuclei as tiny magnets, the nuclear spins can be oriented such that the magnetic dipole is aligned parallel or antiparallel to the external magnetic field. These two orientations are of different energies, and the states are separated by an energy difference ΔE. In general, the orientation parallel to the magnetic field is the state with the lower energy, the antiparallel

orientation is the higher energy state. The splitting of the two spin states into two levels of different energy by an external magnetic field is known as the *Zeeman effect* (Figure 22.1). The Zeeman effect thus generates the two-state system required for NMR spectroscopy that are inherently present in the molecules in other types of spectroscopy: in NMR (and EPR, see Section 22.2) the energy states that are later used for measurements have first to be created by applying a magnetic field.

In the presence of a magnetic field, we can induce transitions between the lower and the higher energy level by irradiating the sample with electromagnetic radiation of the correct energy to provide the energy difference ΔE. The difference between the energy levels depends on the strength of the external magnetic field. In an external field of strength B_0 in the z-direction, the energy difference for a spin ½ nucleus is

$$\Delta E = 2m_z B_0 \qquad \text{eq. 22.1}$$

where m_z is the component of the magnetic moment of the nucleus that is parallel to the external field. The magnetic moment in the z-direction is dependent on the *magnetogyric ratio* γ of the particular type of nucleus according to

$$m_z = \frac{h\gamma}{4\pi} \qquad \text{eq. 22.2}$$

The magnetogyric ratio depends on the charge and the mass of the nucleus (Table 22.1). It is positive for ^{1}H and ^{13}C, but negative for ^{15}N. The sign of the magnetogyric ratio determines which orientation of spins is the lower energy state: for nuclei with positive magnetogyric ratio γ, the orientation parallel to the external field is energetically more favorable, for nuclei with negative γ, the antiparallel orientation is the lower energy state. By combining eq. 22.1 and eq. 22.2, we obtain

$$\Delta E = \frac{h\gamma}{2\pi} B_0 \qquad \text{eq. 22.3}$$

Even with the strongest magnetic fields in commercially available spectrometers today (28.2 T, 1.2 GHz NMR spectrometer), the energy difference ΔE between the two states is tiny compared to the energy differences between electronic or vibrational states. According to the Boltzmann equation (see Section 3.1, Section 21.1.2, eq. 21.7, eq. 21.8), the different energy levels are therefore populated almost equally, with just a slight excess of nuclei in the lower energy state. With stronger magnetic fields, the energy gap and the population difference increase, leading to higher sensitivity of NMR experiments with increasing magnetic fields.

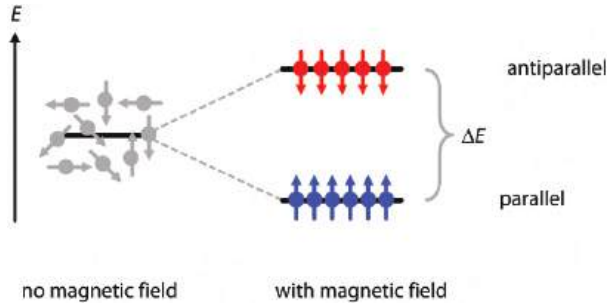


FIGURE 22.1: The Zeeman effect. Nuclei with spin = ½ can exist in two different energetically different states in the presence of an external magnetic field (red, blue), a phenomenon called Zeeman effect. Due to the small energy differences ΔE, the two states are almost equally populated.

With the relation $\Delta E = h\nu$, we can calculate the frequency ν of the electromagnetic radiation that is required to induce transitions between two spin states as

$$\nu = \frac{\gamma}{2\pi} B_0 \qquad \text{eq. 22.4}$$

For a magnetogyric ratio of $268 \cdot 10^6$ T^{-1} s^{-1} (hydrogen, Table 22.1) and typical magnetic field strengths of NMR spectrometers in the 10 T range, we need radiation with a frequency of several 100 MHz (wavelength of about 1 m), *i.e.* radiofrequency, to induce transitions of hydrogen spins between the ground state and the excited state in NMR. This frequency is related to the *Larmor frequency* (Section 22.1.2) of the hydrogen nucleus. In fact, NMR spectrometers are named according to the resonance frequency of protons at their respective magnetic field strengths: a 500 MHz spectrometer has a magnetic field of 11.8 T, which translates into a Larmor frequency of 500 MHz for protons. The 23.5 T magnetic field of a 1 GHz spectrometer leads to a Larmor frequency for protons of 1 GHz.

TABLE 22.1:
NMR parameters of biologically relevant nuclei.

nucleus	spin quantum number	relative abundance (%)	magnetogyric ratio γ (10^6 T^{-1} s^{-1})
^1H	½	> 99.9	268
$^2H = D$	1	0.02	41
^{13}C	½	1.1	67
^{15}N	½	0.37	–27
^{19}F	½	100	25
^{31}P	½	100	108

22.1.2 A one-dimensional NMR spectrum: Larmor frequency, chemical shift, *J*-coupling, and multiplicity

22.1.2.1 The Larmor frequency

The application of the external magnetic field not only causes the generation of a two-state system and the alignment of the nuclear spins with the applied magnetic field. The interaction of the magnetic momentum of the nucleus with the magnetic field also leads to precession of the spin around the axis that is defined by the direction of the magnetic field. The frequency of this precession movement is the *Larmor frequency* ν_0. It depends on the magnetogyric ratio γ of the nucleus and on the strength B_0 of the external magnetic field:

$$\nu_0 = -\frac{\gamma}{2\pi} B_0 \qquad \text{eq. 22.5}$$

This frequency is of the same magnitude as the frequency of electromagnetic radiation that is needed to induce transitions between the two spin orientations (eq. 22.4). The Larmor frequency ν_0 is therefore also called the resonance frequency of the particular nucleus.

22.1.2.2 The local magnetic field and the chemical shift

Nuclei of the same chemical identity have the same magnetogyric ratio γ and the same Larmor frequency. Therefore, we would not be able to distinguish these nuclei within a large molecule, such as different protons in a protein. What makes NMR spectroscopy so powerful is the effect of the chemical environment of the nucleus on the Larmor frequency. The Larmor frequency of a specific nucleus in a larger molecule depends on the local magnetic field that this nucleus experiences. The electrons of the molecule shield its nuclei from the external magnetic field, rendering the local magnetic field slightly smaller than the external field. The effect of the electrons on the local magnetic field B_{local} can be described by the *shielding constant* σ:

$$B_{local} = (1 - \sigma) B_0 \qquad \text{eq. 22.6}$$

Particularly strong effects on the local magnetic field come from electronic ring system currents of aromatic groups. The external magnetic field induces ring currents in aromatic systems, and the current in turn generates an additional local magnetic field (Figure 22.2). This induced field is parallel to the external field outside the ring system, but antiparallel inside the ring system. Hence, nuclei in the plane of the ring on the outside experience a higher local magnetic field – they are deshielded. By contrast, nuclei above or below the plane of the ring system experience a lower local magnetic field – they are shielded from the external field. Paramagnetic ions, such as Fe^{2+}, Fe^{3+}, or Mn^{2+}, that are present in the active sites of many enzymes also strongly affect the local magnetic field. It is this effect of electrons on the local magnetic field that allows distinction of the same type of nuclei by NMR.

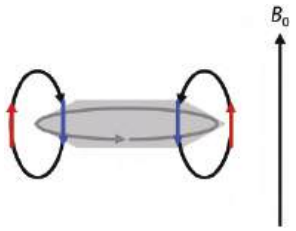


FIGURE 22.2: Local magnetic fields around aromatic ring systems. The external magnetic field (B_0) induces ring currents (gray circular arrow) in the aromatic system (gray hexagon). These ring currents are associated with a magnetic field (black circular arrows) in the direction of the external field (red) outside the aromatic ring, and antiparallel to the external magnetic field inside of the ring (blue). The local magnetic field and the Larmor frequency of a nucleus thus depends on its chemical environment that leads to shielding (blue) or deshielding (red) of nuclei.

As a consequence of the shielding effect of electrons, the same nuclei in different electronic environments within the molecule have slightly different Larmor frequencies. This effect is called the *chemical shift*, because the Larmor frequency is shifted from the expected value (eq. 22.5). The chemical shift δ is expressed in parts per million (ppm) of the unperturbed resonance frequency ν_0 at the external field strength B_0. δ is calculated relative to a reference substance with the Larmor frequency ν_{ref}:

$$\delta = \frac{\nu_s - \nu_{ref}}{\nu_0} \qquad \text{eq. 22.7}$$

where ν_s is the measured Larmor frequency at the local magnetic field strength B_{local}, and ν_0 is the Larmor frequency at the external field strength B_0. While the Larmor frequencies depend on the strength of the external magnetic field (eq. 22.5), the chemical shift δ is independent of the field strength. A common reference substance for proton NMR is tetramethylsilane (TMS). The twelve protons in TMS are in an identical chemical environment and provide a strong reference signal. Shielded nuclei experience a magnetic field $B_{local} < B_0$, and their Larmor frequency is $\nu_s < \nu_0$. The chemical shift δ of shielded nuclei is therefore small. Deshielded atoms experience a magnetic field $B_{local} \approx B_0$, and the Larmor frequency is $\nu_s \approx \nu_0$, which corresponds to larger chemical shifts. In continuous wave NMR experiments where the frequency of the electromagnetic radiation is fixed and the spectrum is measured by changing the external magnetic field B_0 (Section 22.1.2.5), a higher external field B_0 is required for shielded nuclei than in the absence of shielding. The region of the NMR spectrum where resonances from highly shielded nuclei are found is therefore also called the high-field or upfield region. Resonances from deshielded nuclei populate the low-field or downfield region of the spectrum.

It is not possible to assign a particular nucleus to a specific chemical group just on the basis of its chemical shift, but certain types of nuclei show resonances in characteristic regions of an NMR spectrum. Aliphatic protons, for example, typically exhibit low chemical shifts, whereas protons in aromatic systems show higher chemical shifts (Figure 22.3).

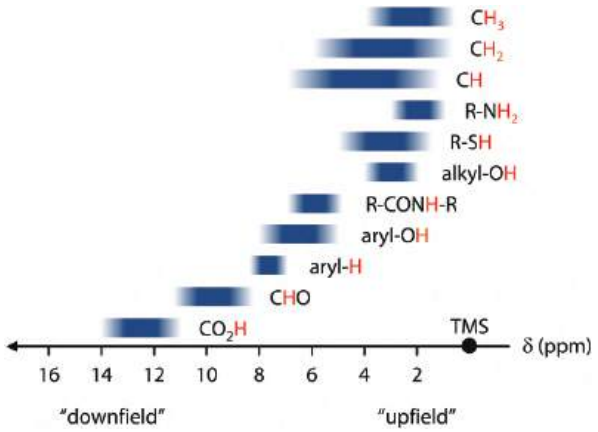


FIGURE 22.3: Chemical shifts for protons in different groups. The chemical shift relative to the reference substance TMS depends on the chemical environment of the proton, resulting in characteristic chemical shifts for protons within functional groups.

Chemical shifts are very sensitive to the local environment and serve as indicators for secondary structure formation in proteins (Box 22.1) or can be used to map interaction sites (Box 22.2).

BOX 22.1: CHEMICAL SHIFTS AND SECONDARY STRUCTURE.

The chemical shift of Cα-protons in proteins is sensitive to the conformation, and depends on the ϕ- and ψ-angles of the polypeptide chain (Section 18.1.2 and Figure 18.21). Cα-protons within α-helices typically show an upfield shift (lower chemical shift) compared to random coil, whereas β-sheet formation is associated with a downfield shift (higher chemical shift), leading to a chemical shift difference $\Delta\delta = \delta_{observed} - \delta_{random\ coil}$. Residues that show $\Delta\delta > 0.1$ are assigned a chemical shift index of 1, whereas $\Delta\delta < -0.1$ corresponds to a chemical shift index of –1. Otherwise, the chemical shift index is zero. Protein regions that adopt α-helical or β-sheet structure can be identified by plotting the *chemical shift index* as a function of the position within the amino acid sequence. At least four consecutive residues with a chemical shift index of –1 indicate α-helix formation. At least three consecutive residues with a chemical shift of 1 point to β-sheet formation. The remaining regions are classified as random coil. Ends of secondary structure elements are marked by a chemical shift index of opposite sign, or by two residues with a chemical shift index of zero. Wishart *et al.* (1992) Biochemistry 31(6): 1647–1651.

BOX 22.2: CHEMICAL SHIFT PERTURBATION TO MAP INTERACTION SITES.

The sensitivity of the chemical shift to the environment can be exploited to map binding sites for ligands in chemical shift perturbation experiments. To map the ligand binding site on a protein, an NMR spectrum of the free protein is measured and the resonances of the amide protons are assigned. In a second step, the NMR spectrum is recorded in the presence of the ligand. Some amide proton resonances will undergo a large change in chemical shift upon ligand binding, some show smaller changes, and some do not change at all. In NMR titrations, chemical shifts are plotted as a function of the concentration of the binding partner to determine the K_d value of the complex. This procedure is possible only when binding and dissociation of the ligand are rapid, and the measured chemical shift is a population-weighted average of the chemical shifts of free and bound states (fast-exchange

(*Continued*)

regime, see Section 22.1.7). Mapping the residues whose amide protons exhibit a large chemical shift perturbation on the structure of the protein reveals the relative location of these residues. A clustering of these amino acids often reveals the binding site for the ligand. However, it has to be noted that chemical shift perturbation may also result from conformational changes. A large chemical shift perturbation therefore cannot directly be taken as a proof for the contribution of a particular amino acid in binding.

22.1.2.3 Scalar coupling and multiplets

A single nucleus, *e.g.* a proton, yields one defined resonance in an NMR spectrum, with the spectral position depending on its electronic environment. In a larger molecule, this nucleus (nucleus 1) is bonded to other nuclei by bonding electrons (Figure 22.4), and these nuclei affect the local magnetic field experienced by nucleus 1. The spin of a bonded nucleus 2 can assume two different orientations. Depending on the orientation of spin 2, nucleus 1 has two slightly different Larmor frequencies, and two resonances for nucleus 1 will be detected in an NMR experiment. This phenomenon is called *scalar* or *J-coupling*. The split resonance resulting from *J*-coupling is called a *doublet*. The separation of the two resonances (measured in Hz) is called the *J-coupling constant*.

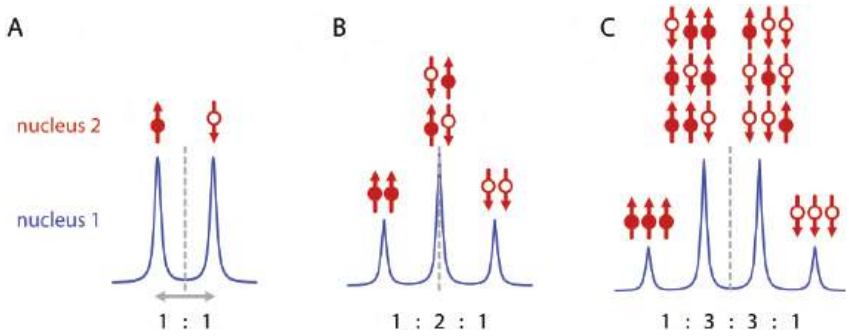


FIGURE 22.4: Interaction between two nuclei by scalar coupling. A: When two nuclei 1 and 2 are connected by chemical bonds, the orientation of the magnetic moment of nucleus 2 (red) is experienced by nucleus 1 (blue), and influences the energy levels of nucleus 1. Coupling of nucleus 1 to nucleus 2 splits the single resonance of nucleus 1 into two resonances of equal intensity, corresponding to transitions of nucleus 1 with the spin of nucleus 2 in either of the two possible orientations (left). The scalar coupling constant *J* is the difference between the frequencies (in Hz) of the two resonances (gray double-headed arrow). B: If nucleus 1 is coupled through chemical bonds to two nuclei 2, the resonance of nucleus 1 is split into a triplet, with the intensities 1:2:1. The orientations of the two nuclei 2 are indicated above the lines of the triplet. C: Coupling to three nuclei 2 splits the resonance of nucleus 1 into a quartet, with an intensity ratio of the four lines of 1:3:3:1, corresponding to the possible orientations of the spins of the three nuclei 2 that are indicated above the line of the quartet (for $\gamma > 0$).

In NMR spectroscopy, *J*-coupling through up to three bonds is typically analyzed. The *Karplus equation* relates the value of *J* for two nuclei coupled through three bonds to the dihedral angle ϕ that is defined by the bonds (see Figure 17.4):

$$J = a\cos^2\phi + b\cos\phi + c \qquad \text{eq. 22.8}$$

The constants *a*, *b*, and *c* depend on the nuclei and the connecting bonds. *J*-coupling constants are between 0.1 and 10 Hz for protons, but can exceed these values for other nuclei. A typical example of *J*-coupling in proton NMR is the splitting of resonances for methyl protons in acetaldehyde into doublets by the neighboring carbonyl proton (Figure 22.5). Note that splitting of the resonance of nucleus 1 into a doublet by nucleus 2 also leads to splitting of the resonance for nucleus 2 by nucleus 1,

with the same *J*-coupling constant. Identical coupling constants for two doublet resonances therefore serve as an indication for which nuclei are coupled, *i.e.* are connected by chemical bonds.

If nucleus 1 has two or three identical neighbors, its resonances are split into a triplet or quadruplet. The area under the resonance lines is proportional to the number of relative spin orientations leading to the particular resonance (Figure 22.6). For acetaldehyde, the resonances of the carbonyl proton are split into a quadruplet due to the four possible orientations of the three spins of the methyl protons (Figure 22.5).

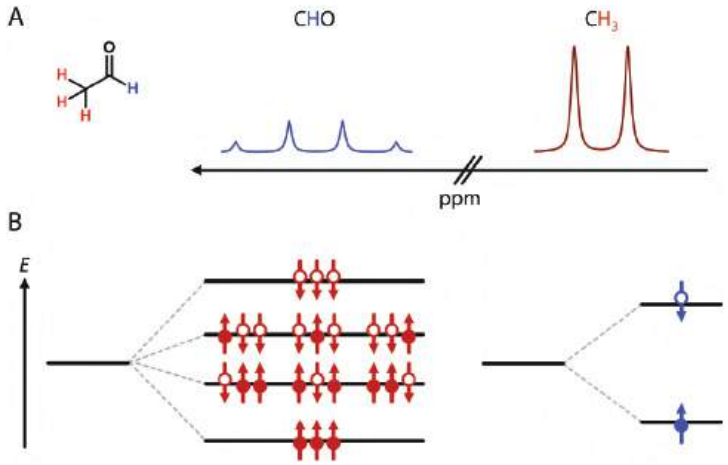


FIGURE 22.5: Scalar coupling in acetaldehyde. A: Acetaldehyde contains one proton on the carbonyl group (blue) and three protons in the methyl group (red). B: The resonance of the three methyl protons is split into a doublet because of coupling with the carbonyl proton whose magnetic moment can adopt two different configurations with different energies. The resonance for the carbonyl proton is split into a quartet because of coupling with the three methyl protons whose magnetic moments can have different combinations of orientation, giving rise to four energy levels.

22.1.2.4 Shape of NMR lines

In emission spectroscopy, resonance lines are *Lorentzian-shaped*. The intensity I as a function of the frequency ω is

$$I(\omega) = I_0 \cdot \frac{\Gamma/2}{(\omega - \omega_0)^2 + (\Gamma/2)^2} \qquad \text{eq. 22.9}$$

I_0 is the amplitude, *i.e.* the intensity at the resonance frequency ω_0, and Γ is the sum of all rate constants that lead away from the excited state. Γ corresponds to the full width at half maximum of the Lorentzian line. For NMR transitions, the relaxation from the excited to the ground state occurs with the *transverse relaxation time* T_2 of spin inversion (see Section 22.1.4). In the absence of other relaxation processes, Γ is

$$\Gamma = \frac{1}{T_2} \qquad \text{eq. 22.10}$$

and we can directly determine the relaxation time T_2 from the width of an NMR line.

Overall, we thus obtain multiple pieces of information from a one-dimensional NMR spectrum (Figure 22.6): the chemical shift is indicative of the chemical environment of a particular proton. The area under the resonance line is proportional to the number of nuclei with the same chemical shift. The line width depends on the relaxation time. The multiplicity of a particular resonance indicates the number of spins with which the respective nucleus shows scalar coupling. The coupling constant is represented by the distance of the individual lines of a multiplet.

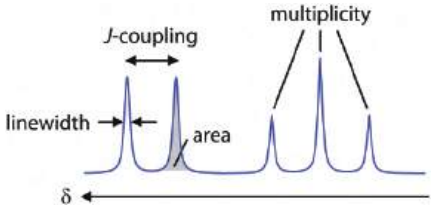


FIGURE 22.6: Information from an NMR spectrum. The chemical shift is characteristic of the local magnetic field experienced by a nucleus. The multiplicity provides information on the number of nuclei that are coupled *via* chemical bonds, and the *J*-coupling is a measure of the coupling strength. The relaxation time (see Section 22.1.4) is encoded in the line width of the resonance, and the area under the resonance or the lines of a multiplet is proportional to the number of nuclei that contribute to this resonance.

22.1.2.5 Instrumentation

A simple NMR spectrometer contains a sample compartment within an electromagnetic or permanent magnet that generates the external magnetic field B_0 (Figure 22.7). This magnetic field establishes the two-state systems by way of the Zeeman effect. A high-frequency radio transmitter provides the radiation for excitation of spins in the sample. When the resonance condition is fulfilled, spin inversion will be detected by a receiver, an electromagnetic coil in which the change in magnetization of the sample induces a current that is measured. Scanning of excitation frequencies can be achieved in two different ways: either, in an approach analogous to wavelength scanning by means of a monochromator in optical spectroscopy, the magnetic field B_0 is kept constant and the frequency ν of the excitation radiation is varied, or the frequency of the excitation is kept constant and the B_0-field is scanned. In this case, the energy differences between states are changed, and the resonance condition for different nuclei is met at different field strengths B_0. Both of these *continuous wave* (cw) methods measure the absorption lines for individual transitions separately, which together give the NMR spectrum.

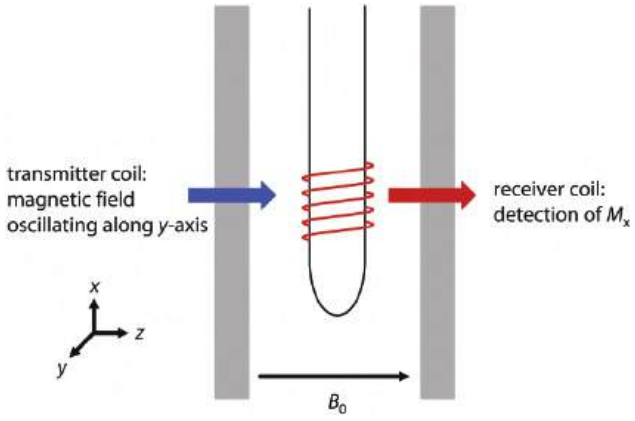


FIGURE 22.7: NMR spectrometer. A static magnetic field B_0 in the *z*-direction is generated by a superconducting magnet (gray) surrounding the sample compartment (black). A transmitter coil (not shown) provides the oscillating magnetic field along the *y*-axis that induces transitions between the two spin states. The oscillating magnetization in the *xy*-plane induces an oscillating current in the receiver coil (red) that is detected as the NMR signal.

22.1.3 The Nuclear Overhauser Effect: Distance Information

J-coupling is mediated by bonding electrons. In contrast, the *nuclear Overhauser effect* (NOE) is caused by an interaction of two spins through space. This dipolar coupling leads to the transfer of magnetization by *cross-relaxation*. Dipolar coupling is strongly distance-dependent: its efficiency depends on the inverse 6th power of the inter-nuclear distance. We have encountered this distance

dependence for the interaction of dipoles before in the distance dependence of London forces (eq. 17.17), and in the distance dependence of Förster resonance energy transfer (eq. 21.131). Because of its strong distance dependence, the NOE provides a powerful tool to obtain intramolecular distance information.

NOEs can be detected by changes in the intensity of a particular resonance, when a second resonance is saturated by continuous irradiation, such that the populations of the two states are equal. The measured intensity I can increase (positive NOE) or decrease (negative NOE) relative to the intensity I_0 without saturation of the second resonance. The nuclear Overhauser enhancement is defined as

$$NOE = \frac{I - I_0}{I_0} \qquad \text{eq. 22.11}$$

To understand the cross-relaxation that leads to the NOE, we consider the energy levels of the two nuclei and the effect of saturation of one transition on the population differences (Figure 22.8). In the unperturbed system of two nuclei A and B, the energetic differences between the energy levels determine the population of the individual states. The lowest-energy state, with both spins in an orientation parallel to the external field (for nuclei with $\gamma > 0$), is the state with the highest population. The two possible states with one nucleus in parallel and the second in antiparallel orientation are isoenergetic and have a lower population. The highest energy level, corresponding to both nuclei in an antiparallel orientation to the external magnetic field, is even less populated. Note that the energy differences between the lowest and the intermediate state and between the intermediate and highest energy level are identical. We assume we have n molecules in the states with intermediate energy, where the spins of A and B have opposite orientations. The Boltzmann distribution tells us that in equilibrium we have $n+x$ molecules in the state with the lowest energy level, and $n-x$ molecules in the highest energy level. Irradiation at the Larmor frequency of nucleus A will now alter the populations. Upon saturation of A, the populations of A in the spin-up and spin-down state (at a constant orientation of spin B) become equal (Figure 22.8). Effectively, the saturation of nucleus A leads to the transfer of $x/2$ molecules from the lower to the higher energy level of each A transition. The populations of the two states that are relevant for the transitions of nucleus B are now $n+x/2-(n-x/2)$. The population difference therefore still equals x, and the intensity of the resonance of B is not affected. The intensity of this resonance will be altered only if the population differences can be reduced by cross-relaxation in two ways: (1) by transitions between the two states of intermediate energy (W_0 *transition*), and (2) by transitions between the highest and lowest energy level (W_2 *transition*). W_0 transitions involve no net spin inversion, whereas W_2 transitions invert both spins. Both W_0 and W_2 transitions are spin forbidden because they involve changes of two spin states. These transitions are not observed directly by NMR, but can occur during relaxation. W_0 and W_2 transitions are possible only if the respective nuclei are closer than ca. 0.5 nm. The outcome depends on which of these cross-relaxation pathways dominates. If W_0 is the dominant (faster) process, the population difference for the B transitions will be decreased from x to $x-a$. As a consequence, the intensity of the resonance line for B is reduced (negative NOE, Figure 22.8). If, on the other hand, the cross-relaxation *via* W_2 is faster, the population difference for B will increase to $x+a$. In this case, the intensity of the resonance of B increases (positive NOE, Figure 22.8).

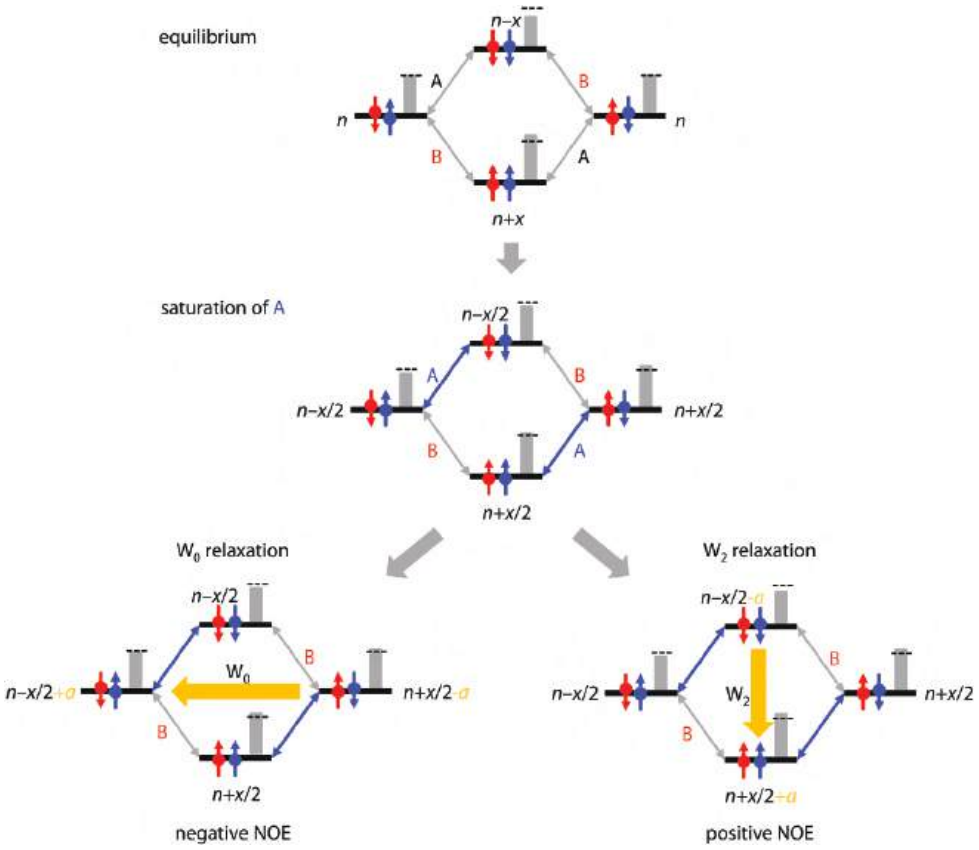


FIGURE 22.8: The nuclear Overhauser effect (NOE). The populations of the possible energy levels of two spins A (blue) and B (red) according to the Boltzmann distribution. We assume that the population of the two intermediate energy levels is *n*. The lower energy level with the two spins parallel to the external field is more populated (*n*+*x*), and the higher energy level with both spins antiparallel to the magnetic field is less populated (*n*–*x*). The gray bars indicate the populations, the dotted black line marks *n*. When we irradiate the sample at the Larmor frequency of nucleus A, the A transitions (red) will be saturated, *i.e.* half of the A-spins in the ground state will be excited, leading to equal populations of the states involved (*n*–*x*/2 or *n*+*x*/2). If the system can now cross-relax by W_0 transitions between the intermediate energy levels, involving a spin flip-flop and no change in overall spin, the population difference for nucleus B will be reduced from *x* to *x*–*a*, and the resonance for B is smaller. This effect is called the negative nuclear Overhauser effect. If, on the other hand, cross-relaxation by W_2 transitions between the highest and the lowest level takes place, and both spins flip orientation in the same directions, the population difference for nucleus B will be increased from *x* to *x*+*a*. In this case, we observe a larger resonance line and a positive nuclear Overhauser effect.

We will see in Section 22.1.6 that the NOE is an important tool for the determination of molecular structures by two-dimensional NMR spectroscopy. The nuclear Overhauser effect can lead to a transfer of magnetization from one molecule to its binding partner. This effect can be exploited to probe binding, and to screen libraries of molecules for binding to pharmaceutical targets (Box 22.3).

BOX 22.3: SATURATION TRANSFER DIFFERENCE-NMR (STD-NMR) TO SCREEN LIGAND LIBRARIES FOR BINDING.

STD-NMR is based on magnetization transfer from a macromolecule to its ligand by the nuclear Overhauser effect. To investigate protein-ligand binding, a resonance of the protein is selectively saturated by electromagnetic radiation at the corresponding Larmor frequency, and the 1D-proton NMR spectrum is measured (on-resonance spectrum). If the ligand binds to the protein, cross-relaxation at the ligand-protein interface allows for transfer of magnetization from the protein to the ligand, which leads to a reduced population difference

(*Continued*)

between the two spin states of the respective proton, and to a reduced intensity in the NMR spectrum. In a control experiment, the saturation is performed with radiation frequency in a spectral region without protein resonances (off-resonance spectrum). The off-resonance spectrum contains resonance lines for all protons in the ligand mixture. The difference spectrum (off-resonance minus on-resonance) then shows the saturation difference, and contains only resonances from ligands to which saturation has been transferred (Figure 22.9).

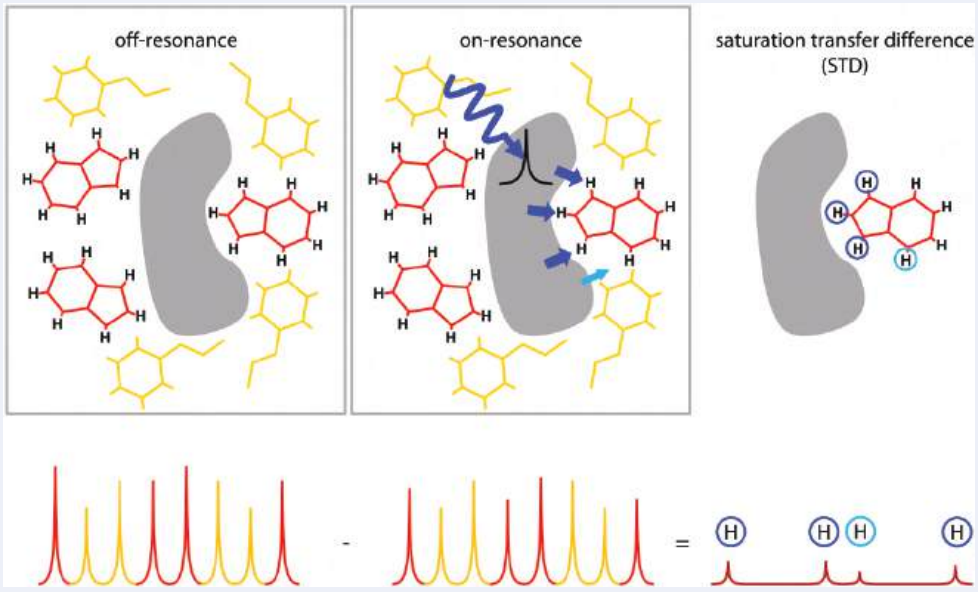


FIGURE 22.9: Principle of STD-NMR. In the off-resonance spectrum, resonances for all protons are recorded. The on-resonance spectrum is taken after selective saturation of a protein resonance. Due to magnetization transfer from the protein to bound ligands, the resonances of protons involved in binding are reduced in intensity. The saturation transfer difference (STD) spectrum shows only resonances from the ligand that binds. The closer the proton to the protein-ligand interface, the higher the intensity of its resonance line.

STD-NMR relies on exchange between bound and free ligand during the experiment. Saturation transfer occurs while the ligand is bound, and the more ligand molecules bind and dissociate from the binding site during the experiment, the higher the transfer difference. Therefore, STD-NMR is ideally suited to detect low-affinity interactions, with K_d values in the low micromolar to the millimolar range. The method can be used for screening and for the identification of lead structures in pharmaceutical research. STD-NMR also provides information on the functional groups of the ligand that mediate binding. Magnetization transfer is achieved through the strongly distance-dependent nuclear Overhauser effect, and saturation is therefore most efficient for ligand protons that are closest to the binding surface. The relative intensity of ligand resonances in the STD spectrum provides information on the orientation of the ligand relative to the binding site and on the protons involved in binding.

22.1.4 Magnetization and its relaxation to equilibrium: Fourier transform-NMR and the free induction decay

In the absence of a magnetic field, the spins of the individual nuclei have random orientations relative to each other, and their magnetic moments cancel over the whole sample. Once an external

magnetic field is applied, however, quantum mechanics restricts the spin of each nucleus to two different orientations. Each individual spin precesses around the *z*-axis (of the external magnetic field), but the precession of the individual spins is not correlated. Consequently, the vector components of the magnetization in the *xy*-plane cancel. In contrast, the magnetization in the *z*-direction does not cancel because the equilibrium population of spins parallel to the external field is slightly higher than the population with an antiparallel orientation. The magnetic moments of all spins therefore add up to a macroscopic *magnetization* in the *z*-direction (Figure 22.10).

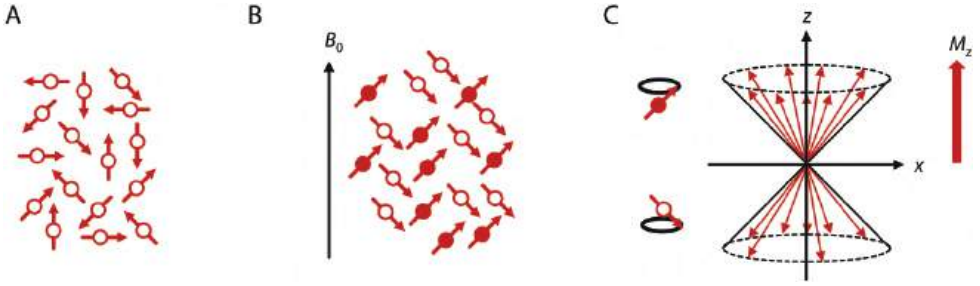


FIGURE 22.10: Generation of macroscopic magnetization. A: In the absence of an external magnetic field, the orientations of the nuclear spins are random, and their magnetic moments (arrows) cancel. B: In the presence of an external magnetic field, the spins are oriented either parallel (filled circle) or antiparallel (open circle) to the B_0-field. C: The tip of the magnetization vector (arrow) precesses around the *z*-axis. The precession of the individual spins is not correlated, and the vector component of the magnetization in the *xy*-plane cancels. Because of the slight excess of spins in the orientation parallel to the external field, their components add up to a magnetization M_z in the *z*-direction.

Now imagine we apply a short pulse of electromagnetic radiation at the Larmor frequency of our nuclei to induce transitions between the two energy levels. For equal populations of both states, the *z*-magnetization will disappear, for population inversion the net magnetization in the *z*-direction will be inverted. What happens to the magnetization in the *xy*-direction? At the beginning of the NMR experiment, before applying the pulse, the magnetization in the *xy*-plane is zero. When transitions between the two states are induced by the radiofrequency pulse, the spins will start to precess in phase. This *phase coherence* comes from the common orientation of the spins when the radiofrequency is applied, and from their precession with the same Larmor frequency. Their in-phase precession leads to a net magnetization in the *xy*-plane that rotates with the Larmor frequency. Overall, the excitation pulse reduces the magnetization in the *z*-direction, and introduces a magnetization in the *xy*-plane. Vector addition of these components gives the overall magnetization *M* that can be regarded as the magnetization in the *z*-direction that has been tilted away from the *z*-axis by an angle α (Figure 22.11). A pulse that causes equal populations of both energy levels will convert the initial macroscopic magnetization in the *z*-direction into magnetization in the *y*-direction. Such a pulse thus tilts the overall magnetization by 90°, and is therefore called a 90° pulse. A 180° pulse leads to population inversion and converts the initial *z*-magnetization into negative magnetization (–*z*). In general, the angle α is related to the duration Δt of the pulse by

$$\alpha = \gamma B_1 \Delta t \qquad \text{eq. 22.12}$$

where B_1 is the magnetic field induced by the electromagnetic radiation of the pulse, and γ is the magnetogyric ratio of the nucleus. When the excitation pulse stops, the system will return to equilibrium by *relaxation* (Figure 22.11). Two processes contribute to this relaxation. On the one hand, the population excess for the parallel orientation is restored by spin transitions, leading to an increase of the *z*-magnetization to its original value. This longitudinal relaxation (along the axis defined by the external field) occurs exponentially with the *longitudinal relaxation time* T_1, the *spin-lattice relaxation time*. On the other hand, the loss of phase coherence between the precessing spins reduces the *xy*-magnetization to zero with the *transverse relaxation time* T_2, the *spin-spin relaxation time*. T_1 and T_2 relaxation times can be measured by defined pulse sequences (Box 22.4).

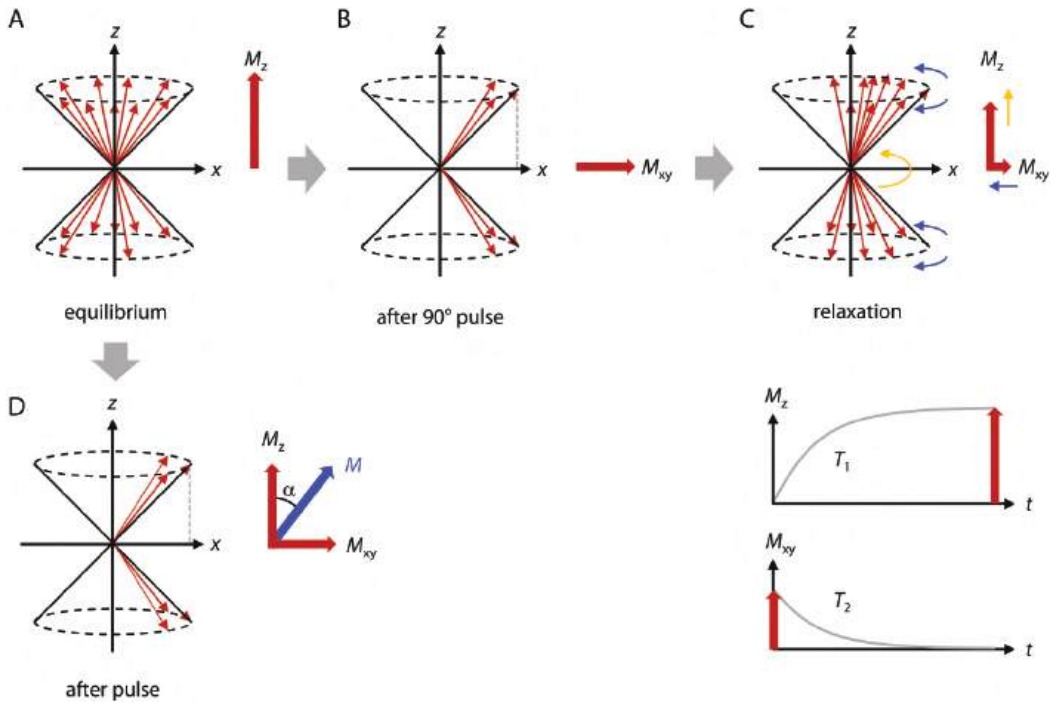


FIGURE 22.11: Magnetization, magnetization flipping, and relaxation. A: In equilibrium, more nuclei have their magnetic dipole oriented parallel to the external magnetic field in the *z*-direction than anti-parallel ($\gamma > 0$). The excess of spins with parallel orientation causes a net magnetization in the *z*-direction, M_z. B: Excitation leads to the inversion of spins into the antiparallel orientation of the magnetic moments, and a decrease of M_z. When both orientations are equally populated, M_z is zero. At the same time, the precession of spins around the *z*-axis is synchronized: they precess in phase. The correlation of spin precession leads to a net magnetization in the *xy*-plane, M_{xy}. C: Relaxation to equilibrium occurs by two distinct processes. The inversion of spin orientations (orange arrow) restores the equilibrium population and the original M_z magnetization with a time constant T_1, the longitudinal or spin-lattice relaxation time. At the same time, the coherence in precession around the *z*-axis is slowly lost (blue arrows), such that M_{xy} decreases with the time constant T_2, the transverse or spin-spin relaxation time. D: In general, a pulse induces transitions between the two spin states, leading to a reduced population difference and phase coherence. The magnetic moment M_z is reduced, and a magnetic moment appears in the *y*-direction. The overall magnetization *M* is thus tilted by an angle α.

We can measure our NMR signal after pulsed excitation by detecting the magnetization in the *xy*-plane as a function of time with a detection coil. The rotating magnetization will induce a sinusoidal current in the coil that oscillates with the Larmor frequency of the nucleus. This sinusoidal current decays to zero with the relaxation time T_2. The exponentially decaying sinusoidal oscillation is called the *free induction decay* (FID).

BOX 22.4: MEASUREMENT OF SPIN-LATTICE (T_1) AND SPIN-SPIN (T_2) RELAXATION TIMES.

T_1 relaxation times can be measured in *inversion-recovery* experiments (Figure 22.12). First, the magnetization M_z is inverted to $-M_z$ by a 180° pulse. After time *t*, the magnetization is reduced because of relaxation by spin inversion. To measure the size of the magnetization, a 90° pulse is applied that brings the magnetization into the *xy*-plane where it can be detected. A series of measurements with different values for *t* gives a set of NMR spectra from which the peak height for each resonance can be determined as a function of *t*. From the time constant τ derived from the exponential dependence of the peak height on *t*, the individual T_1 relaxation time for each resonance can be determined. T_2 relaxation times are

(*Continued*)

typically measured in spin-echo experiments. The M_z magnetization is brought into the *xy*-plane by a 90° pulse. Transverse relaxation and the loss of phase coherence contribute to relaxation during the time *t*. Phase coherence is lost because of inhomogeneities of the external magnetic field, and chemically identical nuclei precess at slightly different Larmor frequencies. As a consequence, the magnetization vectors for individual spins spread out, with more slowly precessing vectors at the end, more rapidly precessing ones at the front of the set of vectors. After time *t*, a 180° pulse along the *y*-axis inverts the spins. Now their order is reversed: the slower ones are at the front, the faster ones at the end. Therefore, the faster precessing vectors can "catch up" with the others, leading to a refocusing of the M_{xy} magnetization after the time 2τ. This effect is called the *spin echo*. The spin echo cannot compensate for the transverse relaxation that has taken place, and the spin echo will be smaller than the original magnetization. The decay of the spin echo with time provides the relaxation time T_2. Relaxation times are measured to obtain information on the dynamics of molecules (Section 22.1.7).

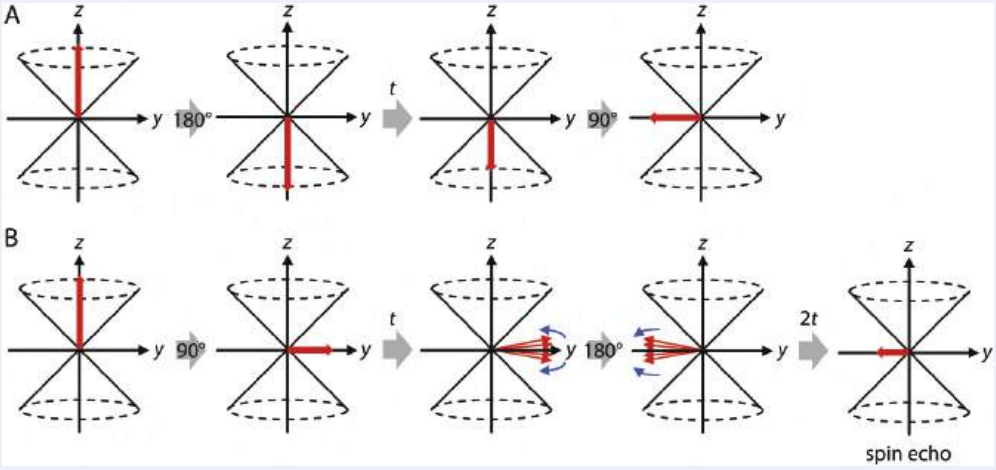


FIGURE 22.12: Measurement of relaxation times. A: Measurement of T_1 by inversion-recovery. B: Measurement of T_2 by spin-echo experiments.

For one type of nucleus with one Larmor frequency, the FID only contains one component. For several nuclei with different Larmor frequencies, however, the FID is the sum of all individual FIDs for each nucleus with its particular Larmor frequency and transverse relaxation time T_2. We can extract the different resonances and their Larmor frequencies encoded in the FID by the mathematical operation called *Fourier transformation* (Section 31.15) that converts the measured signal as function of time into a signal as a function of inverse time = frequency (see also Fourier transformations in IR-spectroscopy, Section 21.4.3, and in X-ray crystallography, Chapter 24). The Fourier transform of the measured magnetization as a function of time is the intensity as a function of frequency, which is our Fourier transform NMR (FT-NMR) spectrum (Figure 22.13).

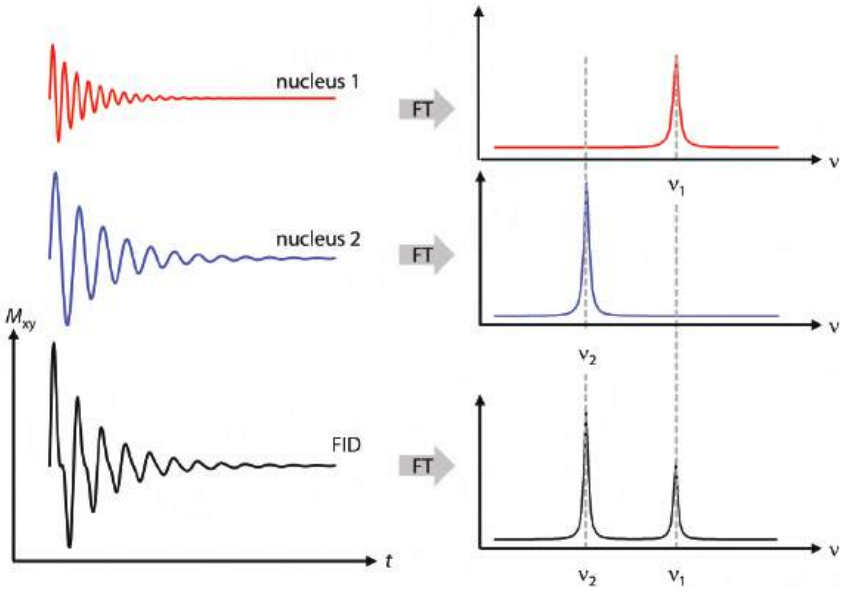


FIGURE 22.13: The FID and Larmor frequencies of nuclei. The nuclei 1 and 2 with their specific Larmor frequencies ν_1 and ν_2 cause a sinusoidal oscillation with the Larmor frequency that decays exponentially with the transverse relaxation time T_2 (red, blue). Fourier transformation (FT) of the exponentially decaying sinusoidal oscillation gives a Lorentzian line at the oscillation frequency ν_1 or ν_2. The measured FID (M_{xy} as a function of time, black) is the sum of the individual FIDs for the two nuclei. Fourier transformation gives the two Lorentzian lines at the Larmor frequencies ν_1 and ν_2. The amplitude of the two individual FIDs determines the intensity of the lines.

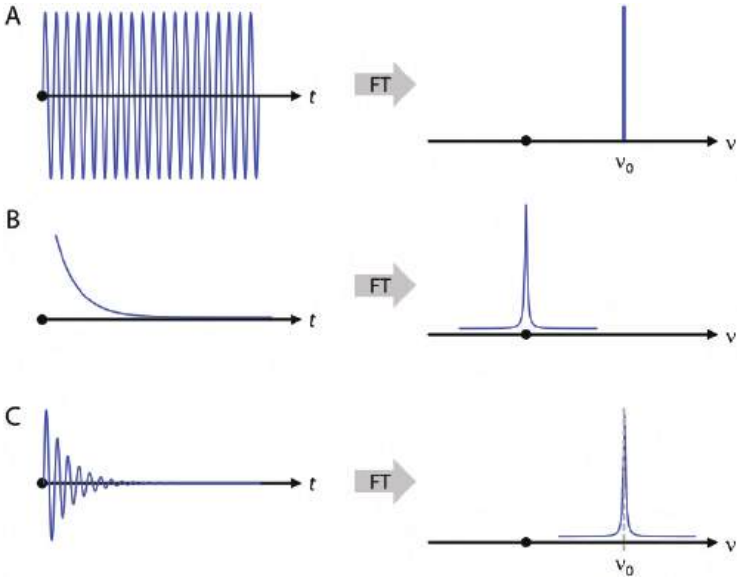


FIGURE 22.14: Fourier transformation of functions describing the FID. A: The Fourier transform of an infinite sinusoidal oscillation is a δ-function at the frequency ν_0. B: Fourier transformation of an exponential function gives a Lorentzian line centered at zero (marked by the black circle on the axis). C: The Fourier transform of an exponentially decaying sinusoidal oscillation, such as the FID, is a Lorentzian line centered at the frequency ν_0.

One important feature we have not yet discussed is the consequence of using a short excitation pulse. By Fourier transformation of this pulse (intensity as a function of time), we can also analyze the excitation pulse for its frequency components (intensity as a function of frequency). The excitation pulse consists of an infinite sine wave, multiplied by a square function. The Fourier transform

(FT) of the infinite sine wave is a δ-function (Figure 22.14) that has a value of 1 at the frequency f of the wave, but is zero at all other x-values. The Fourier transform of a square function is a $(\sin x)/x$ function symmetric to the y-axis. The Fourier transform of the product of the two functions is the convolution (see Box 21.16, Section 31.16) of the two Fourier transformed functions, *i.e.* a $(\sin x)/x$ function (Figure 22.15), shifted along the x-axis such that it is centered at the frequency ν. Thus, the excitation with a short pulse provides different frequencies simultaneously. The shorter the pulse, the broader the spectrum of frequencies it contains. With sufficiently short excitation pulses, we can therefore induce transitions of all nuclei of interest simultaneously. NMR using short excitation pulses and measuring the FID is called *Fourier transform-NMR* (FT-NMR).

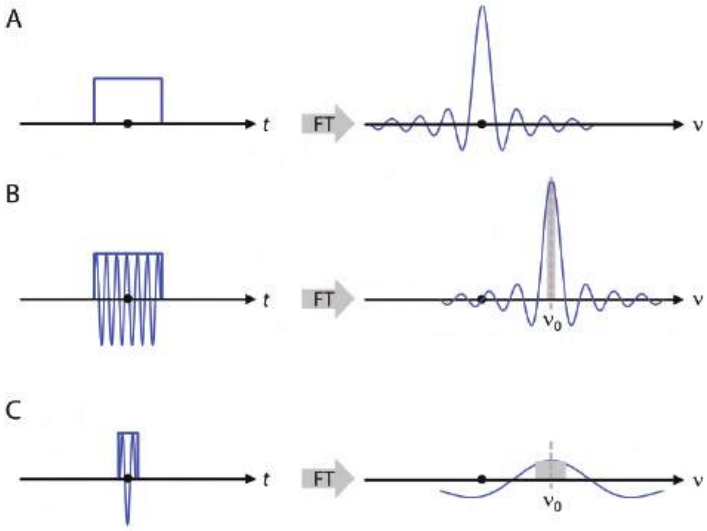


FIGURE 22.15: Fourier transformation of functions describing the excitation pulse. A: The Fourier transformation of a rectangular pulse is a $(\sin x)/x$ function, centered at zero. The shorter the pulse, the broader the $(\sin x)/x$ function. B: Fourier transformation of a short pulse of a sinusoidally oscillating intensity gives a $(\sin x)/x$ function, centered at the frequency ν_0 of the sine function. C: Shorter pulses correspond to a wider $(\sin x)/x$ function, and can provide a wider range of frequencies around ν_0 (gray).

22.1.5 Two-dimensional FT-NMR: COSY and NOESY

NMR spectra of larger molecules with increasing number of nuclei become more and more complex. It is therefore increasingly difficult to assign resonances to a particular nucleus, and to directly determine the structure of the molecule of interest. At some point, the spectrum will be so busy that resonances overlap and can no longer be distinguished. *Two-dimensional (2D) NMR* introduces a second frequency axis, and thereby enables separation of overlapping resonances along both axes. A conceptual analogy is the separation of proteins by gel electrophoresis in two dimensions. In the first dimension the proteins are separated according to their molecular mass. However, proteins with similar molecular mass will not be separated, and their bands will overlap. We now perform a separation in a second dimension perpendicular to the first. This time we separate not by the molecular mass but by a different parameter, such as the isoelectric point. If the proteins with similar molecular masses have different isoelectric points, they will be separated in the second dimension into non-overlapping spots. We will now first introduce the general concept of 2D-NMR, and then see how 2D-NMR can be employed to determine high-resolution structures of proteins and nucleic acids.

22.1.5.1 Principle of a 2D-FT-NMR experiment

Imagine all spins in our sample are in equilibrium, and we start our FT-NMR experiment by applying a 90° pulse, rotating the z-magnetization onto the y-axis. For a certain time interval t_1, we now let the system evolve. During this evolution time, the spins will precess with their Larmor frequencies

around the z-axis. As a consequence, the macroscopic magnetization will rotate in the xy-plane with the Larmor frequency, leading to a decrease in y- and a concomitant increase in x-magnetization. At the same time, transitions of spins between the two states will increase (and eventually restore) the z-magnetization. After the time t_1, we apply a second 90° pulse that rotates the y-component of the magnetization into the $-z$-direction, and then measure the FID of the x-magnetization (Figure 22.16). After all spins have relaxed to equilibrium, we repeat this experiment with a different evolution time t_1. The series of FIDs with different t_1-times can then be converted into frequency space by Fourier transformation. Thus, in 2D-NMR, we perform Fourier transformations in two dimensions, one dimension with the evolution time t_1 and in the second dimension with the measurement time t_2 of the FIDs.

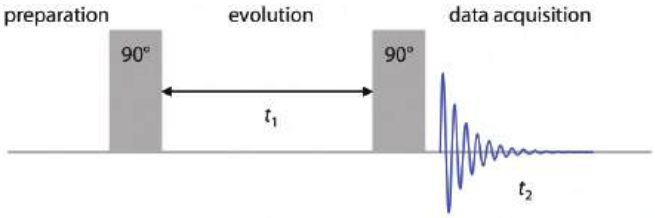


FIGURE 22.16: Principle of a 2D-NMR experiment. During the preparation phase, the equilibrium between ground state (spin parallel to the external field) and excited state (antiparallel) is established, leading to a net magnetization M_z. The 90° pulse rotates the magnetization onto the y-axis. During the evolution time t_1, phase coherence can be transferred between nuclei. Longitudinal relaxation restores the z-magnetization, and transverse relaxation reduces the magnetization in the xy-plane. After t_1, a second 90° pulse is applied, and the FID is measured as a function of the time t_2.

After the first 90° pulse the in-phase precession of individual spins in the same chemical environment gives rise to the magnetization rotating in the xy-plane. The phase coherence of these spins can be transferred to a second group of spins. Phase coherence transfer is in principle possible between nuclei that interact, either by scalar J-coupling through bonds or by dipolar coupling through space by the nuclear Overhauser effect (NOE, Section 22.1.3). Phase coherence transfer is detected through a process called frequency labeling. For a group of spins in the same chemical environment, the entire magnetization will be in the y-direction after the first 90° pulse at $t_1 = 0$, with no component along the x-axis. Due to the precession of the spins and the rotation of the magnetization in the xy-plane with the Larmor frequency ν, the component of the magnetization along the y-axis will then decrease, and the component along the x-axis will increase. We can calculate the magnetization in the x- and y-directions, M_x and M_y, at t_1 as

$$M_x = M_{xy}\sin(2\pi\nu t_1) \qquad \text{eq. 22.13}$$

and

$$M_y = M_{xy}\cos(2\pi\nu t_1) \qquad \text{eq. 22.14}$$

When we apply the second 90° pulse at t_1, the y-component of the magnetization ends up in the $-z$-direction. This makes the y-component undetectable, but the x-component is not affected by the second 90° pulse and can be detected as the measured signal. This magnetization depends on the Larmor frequency ν and the time t_1 according to eq. 22.14. The measured signal oscillates sinusoidally, "labeled" with the frequency ν (Figure 22.17).

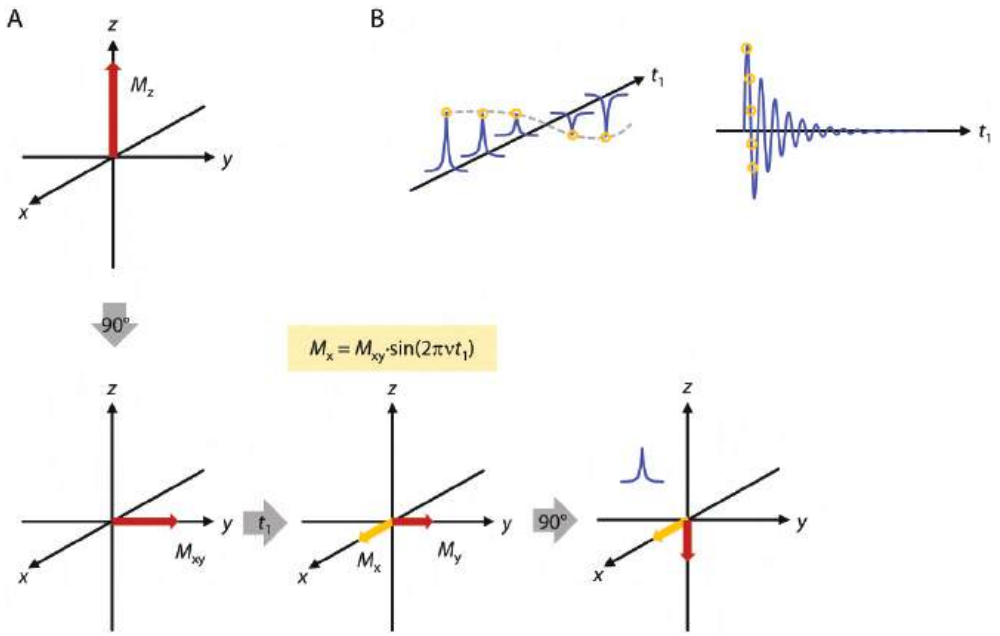


FIGURE 22.17: Magnetization and measured signal during a 2D-NMR experiment. A: *z*-magnetization in equilibrium is turned onto the *y*-axis by a 90° pulse. After the evolution time t_1, the *y*-component is reduced to M_{xy} cos2πνt_1, the *x*-component M_x is M_{xy} sin 2πνt_1. A second 90° pulse converts the *y*-magnetization into the –*z*-direction. The *x*-magnetization is not affected by the detection pulse and is detected. B: The measured signal oscillates sinusoidally as a function of t_1 and decays with the transverse relaxation time T_2. Thus, the signal as a function of t_1 also looks like an FID.

In 2D-NMR spectroscopy, we thus measure a series of FIDs as a function of t_2 with different evolution times t_1. The signal along both time axes oscillates sinusoidally. Fourier transformation along t_2 generates a series of 1D spectra with resonances at the corresponding Larmor frequencies ν. The intensity of each signal in these spectra is modulated sinusoidally and carries the label of the frequency ν. Fourier transformation along t_1 also gives a 1D spectrum with a signal at the frequency ν. Two-dimensional Fourier transformation along t_1 and t_2 generates the 2D-NMR spectrum. For one type of nucleus, the resulting signal is one single resonance at the coordinates (ν,ν). If we now manage to transfer phase coherence from one group of nuclei that precesses with the Larmor frequency ν_1 during t_1 to a second group of nuclei with the Larmor frequency ν_2 that we detect during our acquisition time t_2, the measured signal at the frequency ν_2 is labeled with the frequency ν_1. The resulting peak in the 2D-NMR spectrum obtained by two-dimensional Fourier transformation is located at the coordinates (ν_1,ν_2). This *cross-peak* at (ν_1,ν_2) tells us which groups of nuclei are coupled. Coupling can be observed between nuclei of the same element (*homonuclear coupling*) or between nuclei from different elements (*heteronuclear coupling*).

22.1.5.2 Correlated spectroscopy

We now consider a two-spin system with two spins X and Y that show scalar coupling. The system has four possible energy states, depending on the state of the individual spins (Figure 22.18). In a one-dimensional NMR spectrum, we observe two doublets of resonances, one for spin X, one for spin Y. In *two-dimensional correlated spectroscopy* (COSY), we apply an initial pulse that excites all resonances, in this case X and Y. A second pulse applied after the evolution time t_1 then leads to sharing of phase coherence between coupled spins. We detect the FID with the measurement time t_2, and repeat the experiment with a different evolution time t_1. The pulse sequence for a COSY experiment is 90°—t_1—90°—t_2.

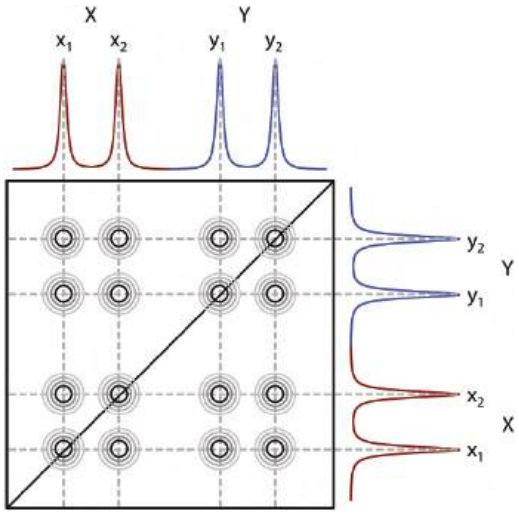


FIGURE 22.18: COSY spectrum of two coupled nuclei. Scalar coupling between the nuclei X and Y leads to splitting of their resonances into doublets (red, x_1 and x_2; blue, y_1 and y_2). In a two-dimensional spectrum, scalar coupling between nuclei gives rise to cross-peaks (off-diagonal concentric circles). The resonances on the diagonal correspond to the 1D spectrum.

By two-dimensional FT, we obtain the 2D-COSY spectrum that contains 16 peaks: for precession with the Larmor frequency of x_1 during t_1 and of y_1 during t_2, of x_2 during t_1 and of y_1 during t_2, of y_1 during t_1 and of x_1 during t_2, of y_1 during t_1 and of y_2 during t_2, and so on (Figure 22.18). From the presence of a cross-peak in a COSY spectrum at (ν_1,ν_2), we can thus conclude that there are nuclei with resonance frequencies ν_1 and ν_2 whose spins interact by scalar coupling.

In the absence of coupling, we would only observe four peaks, namely x_1/x_1, x_2/x_2, y_1/y_1, and y_2/y_2. These are the peaks on the diagonal of the 2D spectrum. The diagonal of the 2D spectrum corresponds to the spectrum we would measure by one-dimensional NMR.

22.1.5.3 Nuclear Overhauser Enhancement Spectroscopy

We can also use 2D-NMR to detect coupling of spins through space, making use of the nuclear Overhauser effect (Section 22.1.3). In a COSY experiment, we apply the second pulse to transfer phase coherence. This pulse brings the magnetization in the y-direction into the $-z$-direction, such that it cannot be detected. In *Nuclear Overhauser Enhancement Spectroscopy* (NOESY), we apply a third pulse after the *mixing time* t_m. This pulse brings the magnetization from the $-z$-direction back into the xy-plane, and we can detect it. The pulse sequence for a NOESY experiment is $90°—t_1—90°—t_m—90°—t_2$.

We now consider two nuclei close in space whose spins interact *via* dipolar coupling during the mixing time t_m. Because of cross-relaxation (see Section 22.1.3) during the mixing time t_m, spin X will be detected at the Larmor frequency of spin Y. After Fourier transformation, this coupling gives rise to a cross-peak at (ν_1,ν_2) in the 2D NOESY spectrum.

NOESY spectra provide extremely useful information for structure determination because they originate from the strongly distance-dependent dipolar interaction (Section 17.2.2 and Figure 17.6). NOEs can be grouped into strong, intermediate, and weak, corresponding to distances of the interacting nuclei of 0.18–0.25 nm, 0.18–0.35 nm, and 0.18–0.5 nm. The same lower boundary takes into account that the absence of an NOE cannot necessarily be interpreted as a large distance between nuclei.

22.1.5.4 Spin systems and sequential assignment of protein NMR spectra

In the following, the general procedure for protein structure determination by homonuclear 2D-FT-NMR will be outlined, starting from a 2D-COSY and a 2D-NOESY ^{1}H spectrum of the protein of interest. These spectra show numerous peaks along the diagonal, corresponding to the 1D spectrum, plus a number of cross-peaks outside the diagonal. The cross-peaks in the COSY spectrum provide information on scalar coupling, *i.e.* on protons that are connected through chemical bonds. The first step is to assign these resonances to individual protons. In light of the number of protons this task seems

rather demanding, if not impossible. The assignment is simplified, however, by the limited number of building blocks in proteins. The important point to note is that detection of scalar coupling is limited to nuclei separated by up to three bonds. Scalar coupling does not extend over the peptide bond, which means that there is no scalar coupling between the Cα proton of an amino acid and the NH proton of the next amino acid. Scalar coupling is therefore limited to protons of the same amino acid, which constitute a *spin system*. Each type of amino acid gives a characteristic coupling pattern that is determined by its chemical structure. Gly has the simplest spin system with only one proton resonance from its Cα protons. Ala gives rise to two resonances, one for the Cα proton and one for the three Cβ protons. The Cα proton and the Cβ protons are separated by three bonds and show scalar coupling, leading to a cross-peak in the COSY spectrum (Figure 22.19). Each Ala in the protein will give rise to a set of these resonances, shifted slightly relative to each other because of different chemical environments. Similarly, each Val, Leu, etc. constitutes a spin system and gives rise to a characteristic peak pattern (Figure 22.19).

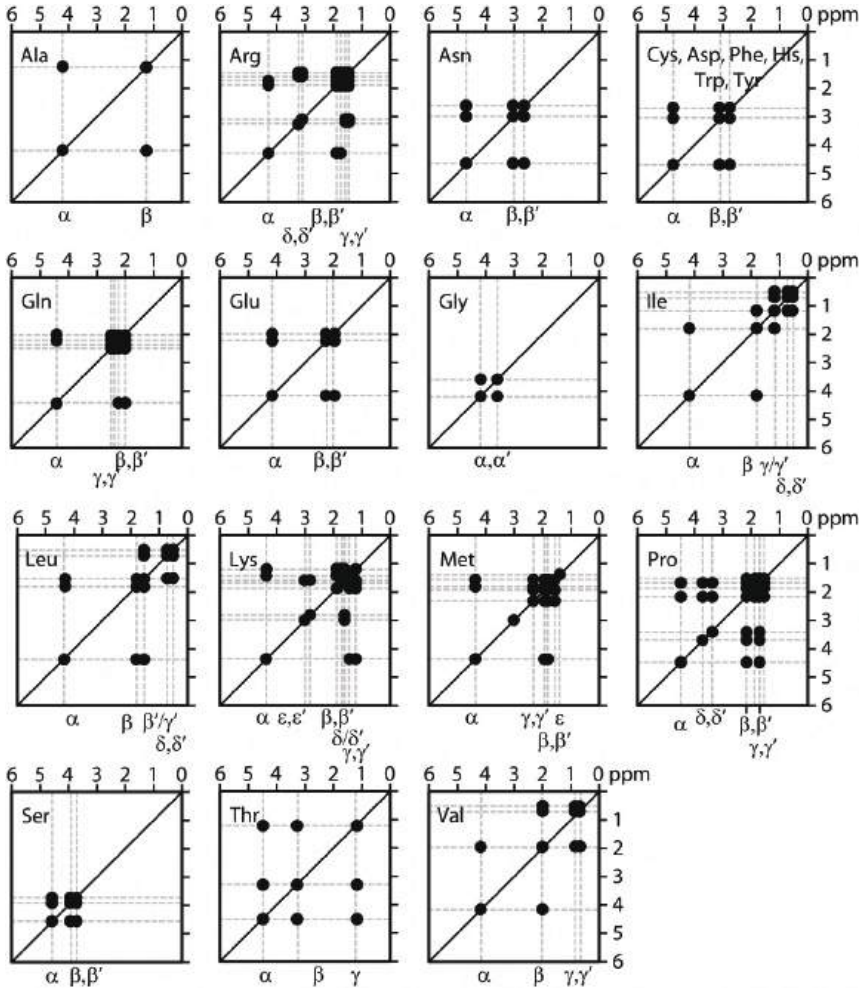


FIGURE 22.19: Spin systems of the 20 proteinogenic amino acids. The chemical structure of amino acid side chains determines the scalar coupling pattern of their protons. The expected COSY cross-peak pattern for the 20 proteinogenic amino acids is shown. Note that Cys, Asp, Phe, His, Tyr, and Trp show the same cross-peak pattern, but the individual values for the chemical shift of their protons are slightly different.

First, we have to identify the different spin systems and assign their resonances to the individual amino acids in our protein. In a second step, we have to find out the exact positions in the protein sequence to which the spin systems belong. This information is not encoded in the COSY spectrum because scalar coupling does not cross the peptide bond. To identify amino acids that are sequential neighbors, we have to resort to NOESY spectra for what is called the *sequential assignment* (Figure 22.20). Cross-peaks from the NOESY spectrum give information on which protons are closer than

0.5 nm. The most important cross-peak for sequential assignment is the NOE cross-peak between the Cα of one amino acid and the N-H proton of the subsequent amino acid. Using this resonance, we can cover the gap from one amino acid to the next. The area of the NOESY spectrum where these cross-peaks are found is called the fingerprint region. We start with the Cα proton resonance for one amino acid (identified in the COSY spectrum) and look for cross-peaks to N-H protons in the NOESY spectrum. Information on the protein sequence is useful for this "chain tracing" because it tells us which spin systems should be connected. If we have an Ala-Val pair in the sequence, we have to look for Cα protons from an Ala spin system coupled to the N-H resonance of a Val spin system. Should we come to the conclusion that the resonance we have assigned to the Cα of an Ala is coupled to the N-H resonance of a Tyr, but we have no Ala-Tyr pair in our protein of interest, we know that we better check our assignment again. Prolines cannot be assigned sequentially this way because they lack a Cα proton. Hence, the sequential assignment breaks off at each proline in the sequence, and we have to restart tracing at the amide proton of the amino acid that follows the Pro in the sequence.

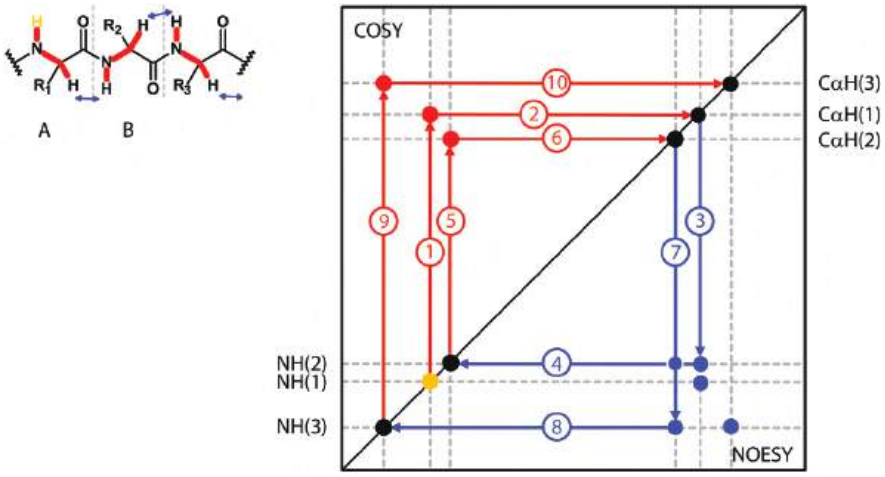


FIGURE 22.20: Chain tracing. Starting at the amide proton resonance for one amino acid (A, orange), the corresponding cross-peak with the Cα proton of this amino acid is identified in the COSY spectrum (1), and traced to the diagonal (2). In the NOESY spectrum (bottom half), the cross-peak between this Cα-proton resonance and the resonance for the N-H proton of the subsequent amino acid B is identified (3) and traced to the diagonal (4). By alternating between COSY and NOESY spectra, the tracing continues (5–10).

At the end of the chain tracing procedure, we have assigned groups of peaks to a spin system pertaining to a specific type of amino acid, resonances within each spin system to individual protons of this amino acid, and each spin system to a particular amino acid in the protein sequence. The remaining NOESY cross-peaks can now directly be assigned to coupling of individual protons from their position in the spectrum (resonance frequencies). NOEs between protons from amino acids distant in sequence provide us with important information on spatial vicinity of these protons in three dimensions. These NOEs are classified as weak, intermediate, or strong NOEs, and provide distance information that is used as conformational constraints for the calculation of the protein structure.

Determination of nucleic acid structures by NMR is more difficult because there are fewer types of building blocks. The nucleobases adenine, cytosine, guanine, and thymine constitute four spin systems. Thymine can be identified by its signature COSY cross-peak between the methyl protons and the C6 proton. In RNA, the number of different spin systems is reduced because C and U show identical coupling. However, both can be separated from purines by their coupling of C5 and C6 protons that is absent in the purine spin systems. The ribose/deoxyribose sugars constitute the fourth/fifth spin system in RNA/DNA. Protons from bases and sugars do not show scalar coupling, but NOESY cross-peaks are found between a base and its own sugar, and to the sugar at the adjacent nucleotide in the 5'-direction. Two principal chain-tracing procedures, based on these NOESY cross-peaks, are possible (Box 22.5). Isotope labeling (deuteration, ^{15}N- or ^{13}C-isotope labeling) can be used to simplify the interpretation of nucleic acid NMR spectra. An additional advantage lies

with the presence of the NMR-active natural isotope ^{31}P, which is directly amenable to NMR studies. In general, the number of long-range distance constraints per building block is smaller than for proteins, and the determination of nucleic acid structures therefore relies more on the correct local geometry. From NOESY spectra, A- and B-DNA (Section 19.5.1) can be distinguished. A-DNA spectra show a high intensity cross-peak for coupling of the C6- (pyrimidine) or C8 (purine) proton and the C2' proton. The cross-peak with the C2'' protons is smaller. In B-DNA, these intensities are inverted because of the different sugar pucker (Section 19.2.1). In addition, A-DNA has a higher helical pitch, and a larger inter-base distance (see Table 19.2). Cross-peaks between the C6/C8 protons of one base and the neighboring base are therefore stronger in B-DNA than in A-DNA.

BOX 22.5: SEQUENTIAL ASSIGNMENT IN ^{1}H NMR OF NUCLEIC ACIDS.

The proton resonances in nucleic acids can be assigned sequentially using NOESY cross-peaks (Figure 22.21). One route uses the cross-peaks between the C6/C8 proton of the base and the H1' of the ribose within the same nucleotide, and between the H1' and the C6/C8 proton of the next base in the 3'-direction. The second route alternates between cross-peaks from the C6/C8 proton of the base to the H2'/2'' of the ribose in the same nucleotide and from the H2'/2'' to the C6/C8 proton of the next base in 3'-direction.

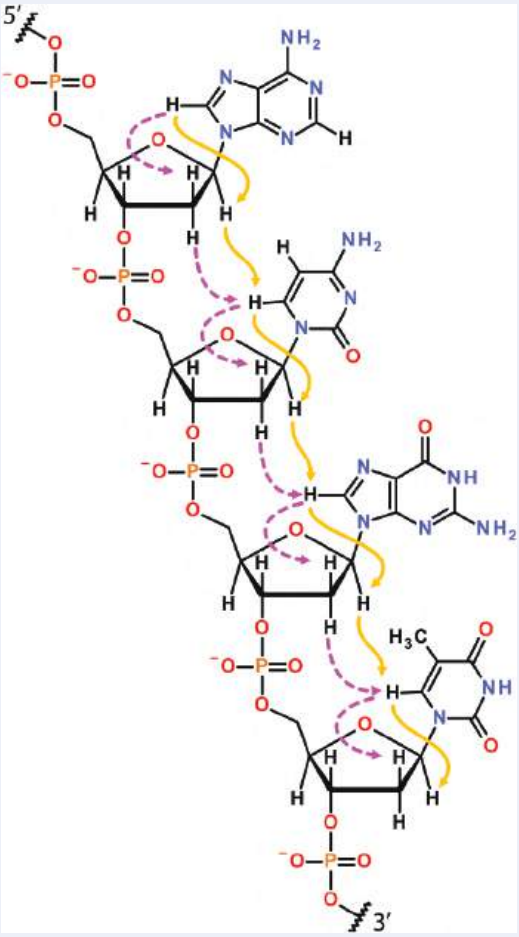


FIGURE 22.21: Sequential assignment of nucleic acid resonances. Orange: sequential assignment by following the cross-peaks from the C6/C8 proton of the base to the C1' proton of the ribose in the same nucleotide, and from the C1' proton of the ribose to the C6/C8 proton of the base in the next nucleotide in the 3'-direction. Magenta: sequential assignment using cross-peaks from the C6/C8 proton of the base to the C2'/2'' protons of the ribose in the same nucleotide, and from the C2'/2'' protons to the C6/C8 proton of the base in the next nucleotide in the 3'-direction.

22.1.5.5 Structure calculation

Finally, we have to convert our NMR restraints into a structural model by expressing all boundary conditions, not only the ones derived from NMR data, in terms of energy. Energy terms express the contributions from covalent and non-covalent interactions (Sections 17.2.1 and 17.2.2). Covalent interactions are specified by bond length, bond angle, and torsion angles for neighboring bonds (eq. 17.3, eq. 17.5, and eq. 17.7). Non-covalent interactions include electrostatic interactions (eq. 17.8), van der Waals interactions (eq. 17.18), and hydrogen bonds. The energy of hydrogen bonds is described by a function similar to the Lennard-Jones potential:

$$E_{\text{Hbond}} = \frac{A'}{r^x} - \frac{B'}{r^y} f(\alpha,\beta) \qquad \text{eq. 22.15}$$

where A' and B' are constants, r is the distance, x and y are empirical parameters, and $f(\alpha,\beta)$ is a function of the angles α and β of the hydrogen atom to donor and acceptor atoms.

In addition to these geometrically defined interaction energies, energy terms reflecting the experimental restraints from NMR such as *J*-coupling or intramolecular distances derived from NOEs are added. The combination of geometric, pre-defined and experimental restraints is the force field (Section 20.1.1), which allows one to calculate the energy of any given structure. Starting from a random coil conformation calculated from the protein sequence, NMR structure calculation is aimed at folding the extended chain into a compact form of minimal energy. Simulated annealing (Section 20.1.3) allows large conformational changes of the starting structure, followed by energy minimization. If an intermediate structure violates the NMR restraints, the corresponding energy terms of the force field apply energetic penalties, and the total energy of the respective structure is high. In this case, additional cycles of minimization are required. The whole process is repeated multiple times until a set of low-energy structures is obtained that fulfill the NMR restraints.

22.1.6 Extending NMR to structure determination of large molecules

NMR of larger proteins becomes difficult because of the increasing number of resonances and the resulting overly crowded spectra. In principle, we can introduce a third (or more) dimension into NMR spectra by introducing a third time period in our measurements. A simple three-dimensional NMR experiment (3D-NMR) then consists of a 90° pulse, followed by an evolution time t_1, a second 90° pulse and a second evolution time t_2, and a third pulse, followed by the measurement time t_3. The NMR spectrum is then obtained by Fourier transformation in three dimensions, leading to a cubic spectrum that can be regarded as layers of 2D-spectra (Figure 22.22). As a result, resonances that overlap or superimpose in a 2D spectrum can now be spread out on the third axis and become separated. Again, 3D-NMR can be performed by transferring magnetization from one type of nucleus to the same type during the evolution times (homonuclear 3D-NMR) or by following the interaction of spins that belong to different types of nuclei (heteronuclear NMR, *e.g.* ^{1}H–^{15}N–^{13}C 3D-NMR). In principle, we can introduce a fourth or fifth dimension into NMR by extending this approach to more evolution times. The number of dimensions in NMR is limited by the relaxation of spins to the ground states and the concomitant loss in net magnetization with time, however.

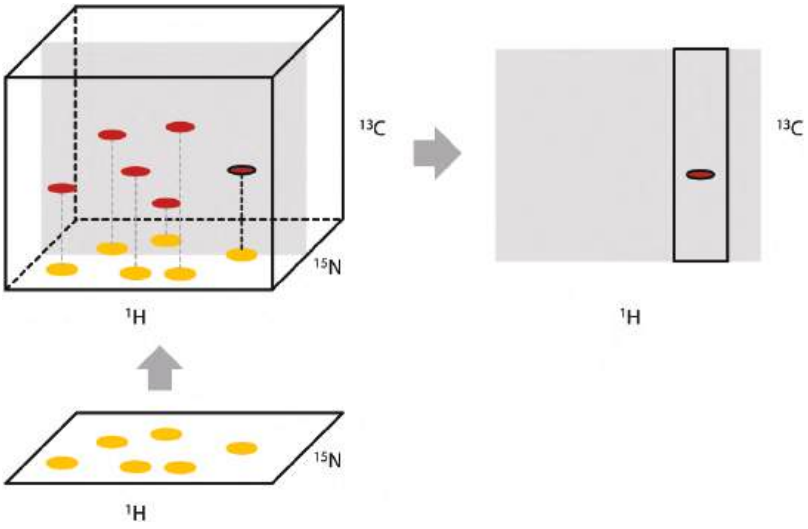


FIGURE 22.22: Principle of 3D-NMR. Special pulse sequences allow detection of coupling between ^{1}H, ^{15}N, and ^{13}C. Cross-peaks in a $^{1}H/^{15}N$ spectrum (orange) are spread (gray dotted line) along the ^{13}C axis according to the coupling with ^{13}C. The 3D-NMR spectrum can be represented as a cube that is formed by a stack of 2D spectra. Coupling between ^{1}H and ^{13}C can be extracted by looking at the layer for a constant ^{15}N chemical shift (gray plane). Often, information from 3D spectra is represented in the form of strips (black rectangle).

We can also limit the number of resonances by *isotope labeling* (Box 22.6). If one type of amino acid is incorporated in isotopically labeled form into our protein of interest, resonances from these amino acids are selectively observed. Alternatively, a segment of the protein can be labeled, and resonances from the labeled area within the protein are selectively observed.

BOX 22.6: ISOTOPE LABELING OF PROTEINS.

Entire proteins can be labeled with ^{15}N and ^{13}C by growing bacteria that overproduce the protein of interest on minimal media supplemented with $^{15}NH_4Cl$ and ^{13}C glucose. Alternatively, only a region of interest, such as a specific domain, can be isotopically labeled. This segmental labeling exploits expressed protein ligation (EPL, Section 21.5.4). One part of the protein is produced in an isotopically labeled form as a C-terminal fusion with an intein. The second part of the protein is produced as an N-terminal intein fusion without isotope labeling. Cleavage of this fusion protein generates this part of the protein with a cysteine at the N-terminus. The cysteine acts as a nucleophile and cleaves the isotopically labeled intein fusion, yielding the protein of interest, segmentally labeled in the N-terminal part. More tailored labeling of only one type of amino acids or specific groups in one amino acid is also possible. In ILV labeling, isoleucine, leucine, and valine are specifically labeled by growing overproducing bacteria on minimal media, providing isotope-labeled forms of the amino acid precursors α-ketobutyrate and α-ketoisovalerate. Methionine methyl labeling results in proteins that are specifically ^{13}C-labeled at their methyl groups. Several other specific labeling schemes have been developed.

In addition to the mere number of resonances causing assignment problems, *line-broadening* is a problem for larger molecules. The line width of NMR resonances is related to the size of the molecule, or more exactly to its rotational correlation time. A small, rapidly rotating molecule gives rise to narrow lines, whereas a larger molecule with a large rotational correlation time shows reduced T_2 relaxation times. Line-broadening increases with the size of the protein until ultimately

the resonances disappear. Two processes are dominant in relaxation, the dipole-dipole coupling and the *chemical shift anisotropy*. While relaxation through chemical shift anisotropy increases with the strength of the magnetic field, relaxation by dipole-dipole coupling is independent of the magnetic field. A special pulse sequence, called *transverse relaxation optimized spectroscopy* (TROSY), has been introduced to achieve cancellation of these two relaxation processes with different signs. Cancellation of these effects can only be achieved at high magnetic field strengths. In TROSY experiments, the T_2 relaxation time is increased, and resolution and sensitivity are increased (Figure 22.23). TROSY may optimally be applied to proteins of > 100 kDa molecular mass.

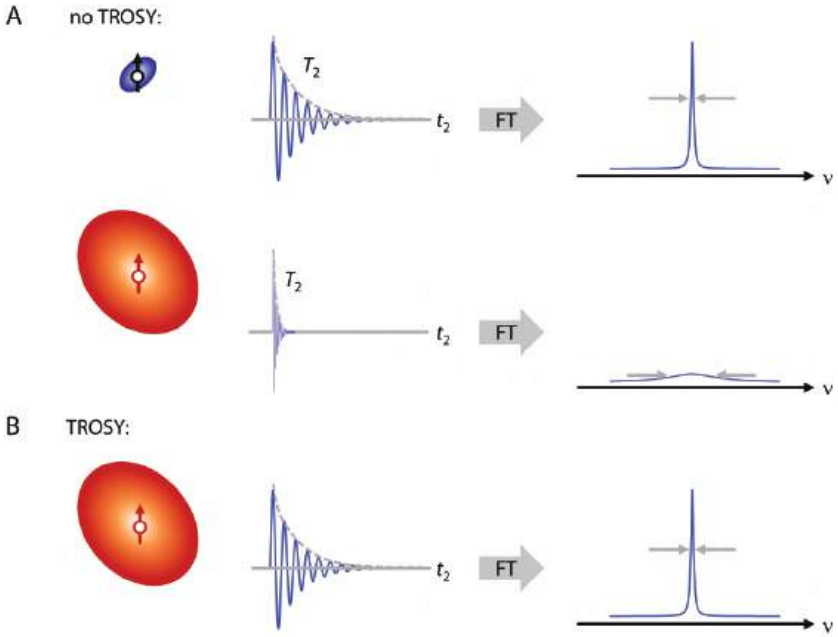


FIGURE 22.23: TROSY. A: The line width of NMR resonances depends on the transverse relaxation time T_2 and on the size of the molecule and its rotational correlation time. Spectra for small molecules consist of narrow lines, whereas the resonances are broadening for larger molecules that tumble more slowly. B: TROSY is a specific pulse sequence that leads to cancellation of two relaxation pathways. As a consequence, T_2 is increased, and the line width is reduced.

22.1.7 NMR AND DYNAMICS

In addition to structural information, NMR can provide information on the dynamics of macromolecules on a large range of time scales, ranging from picoseconds to hours. The method of choice depends on the time scale of the underlying process.

22.1.7.1 Chemical exchange and its effects on NMR resonances

Let us consider a case in which a nucleus exchanges between two different environments, A and B, associated with two different chemical shifts, ω_A and ω_B (Scheme 22.1).

$$\underset{\omega_A}{A} \underset{k_B}{\overset{k_A}{\rightleftharpoons}} \underset{\omega_B}{B}$$

scheme 22.1

The two environments could be caused by the association and dissociation of monomers and dimers, by the presence of two different conformers, or by the presence of free and bound ligand. The effect of chemical exchange on the NMR resonances observed depends on the difference of the resonance frequencies of the nucleus in the two environments, $\omega_A - \omega_B$, and on the rate constant of exchange, k_{ex}. For the reversible reaction in Scheme 22.1, k_{ex} is the sum of the individual rate constants for the forward and reverse reaction, k_A and k_B (see Section 8.1, eq. 8.17 and eq. 8.18):

$$k_{ex} = k_A + k_B \qquad \text{eq. 22.16}$$

The interconversion of these states, or *chemical exchange*, can lead to changes of the resonance frequency, the intensity, and the width of the resonance line. The effects depend on the time scale of the exchange between the two states (Figure 22.24). We have to distinguish three regimes, with fast, intermediate, or slow exchange. If the exchange is very slow ($k_{ex} \ll \omega_A - \omega_B$), then the nuclear spin undergoes many rounds of precession in each environment, with its characteristic resonance frequency, before switching to the other environment. In this case, the nucleus gives rise to two separate signals at ω_A and ω_B. The amplitude of these signals reflects the number of nuclei in the corresponding environments, and hence the equilibrium constant K. If, on the other hand, the exchange is very fast ($k_{ex} \gg \omega_A - \omega_B$), then the time the nucleus spends in each environment is too short to adopt the characteristic precession frequency. Instead, the nuclear spin precesses at an average frequency, which gives rise to a single peak at a resonance frequency between ω_A and ω_B. The position of the resonance line is determined by the relative time the nucleus spends in each state, and hence the equilibrium constant K. With intermediate exchange, in between these extremes, the resonance lines for the two environments coalesce. In this case, the equilibrium constant is difficult to determine.

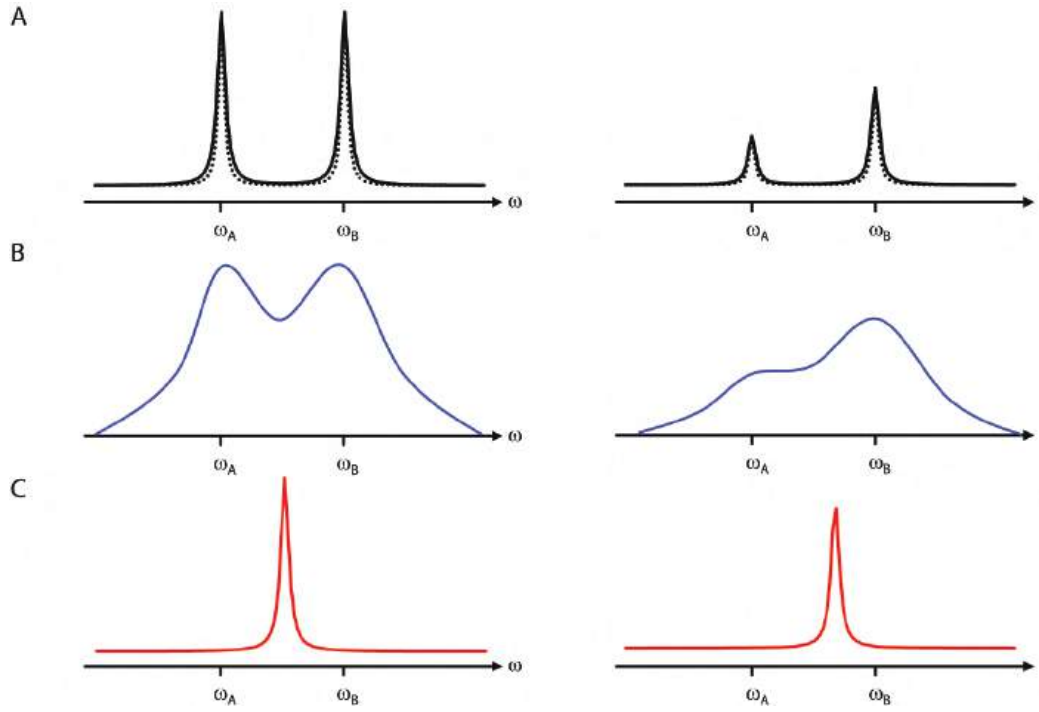


FIGURE 22.24: Chemical exchange and the effect of its time scale on NMR resonances. A: In the slow-exchange regime ($k_{ex} \ll \omega_A - \omega_B$), two resonances are observed at chemical shifts characteristic of environments A and B. The amplitudes provide information on the equilibrium constant. If line broadening is observed, the rate constant for chemical exchange can be determined. The natural line width in the absence of chemical exchange is indicated as a dotted line. B: In the intermediate-exchange regime ($k_{ex} \approx \omega_A - \omega_B$), the two lines for the resonance in A and B coalesce. C: In the rapid-exchange regime ($k_{ex} \gg \omega_A - \omega_B$), one resonance line is detected at a position that corresponds to the population-weighted resonance frequencies associated with environments A and B. The line width is related to the rate constant of chemical exchange. The left panel shows the spectra for equal populations of A and B, *i.e.* for $k_A = k_B$. The right panels show the spectra for a 1:2 ratio of A and B, corresponding to $k_A = 2 \cdot k_B$.

The line width Γ of NMR resonances is determined by the transverse relaxation time T_2 (Section 22.1.4):

$$\Gamma = \frac{1}{T_2} \qquad \text{eq. 22.10}$$

Any process affecting T_2 thus changes the line width of NMR lines. In cases where chemical exchange affects the width Γ of the NMR resonances observed, information on the rate constant of exchange can be extracted by *line shape analysis* (Figure 22.24). Again, the effects of exchange depend on the regime. In the slow-exchange regime, when two individual lines for both states are observed, the boundary condition $k_{ex} \ll \omega_A - \omega_B$ puts an upper limit on k_{ex}. In addition, a large value for either k_A or k_B leads to a depopulation of state A or state B during the time scale of the experiment, which reduces T_2 for this nucleus. The observed line width, Γ_{obs}, then becomes

$$\Gamma_{obs} = \frac{1}{T_2^{obs}} = \frac{1}{T_2^0} + k \qquad \text{eq. 22.17}$$

where T_2^{obs} is the observed transverse relaxation time in the presence of chemical exchange, T_2^0,the transverse relaxation time in the absence of chemical exchange, and k is k_A or k_B. In this case, k_A or k_B can be determined from the analysis of the line width.

In the fast-exchange regime, when $k_{ex} > \omega_A - \omega_B$, rapid averaging over the nuclei in the different environments leads to a single resonance line. Both the resonance frequency and the line width are population-weighted averages of the resonance frequencies and line widths for the two states.

With intermediate exchange, the lines of the two states coalesce into a single, broad line. The line width is increased because the chemical exchange affects the transverse relaxation time T_2. The observed line width is

$$\Gamma_{obs} = \frac{1}{T_2^{obs}} = \frac{1}{T_2^0} + \frac{1}{T_{ex}} \qquad \text{eq. 22.18}$$

where T_{ex} is the relaxation time due to chemical exchange. Here, the exchange does not lead to a loss of the population, as in the slow-exchange regime (eq. 22.17), but causes interference during the time of the measurement. The exchange rate between the two species can be determined from the line width. NMR *line shape analysis* can provide information on the kinetics of binding and dissociation or on conformational changes on the millisecond to second time scale. Line-broadening of small molecule resonances can also be used as a signal for macromolecule binding.

22.1.7.2 Experimental approaches to monitor dynamics

In the slow-exchange regime ($k_{ex} \ll \omega_A - \omega_B$), where the two states cause two different resonances at their characteristic resonance frequencies (Figure 22.24), the conversion of one form to the other can be followed by *real-time NMR*. For this experiment, the reaction is started at t = 0, and the appearance and disappearance of individual or all resonances of the molecule of interest in the initial and final state is then monitored as a function of time. From the time dependence of the signal intensities, the rate constant for the exchange can be determined.

In the slow-exchange regime, structural changes of biomolecules can also be followed through hydrogen exchange. This approach is based on the fact that protons of macromolecules are constantly exchanged with protons of the solvent. The rate of this exchange depends on the exposure of the individual proton to the solvent, and hence the structure of the molecule, as well as the pH. Readily accessible protons on the surface of the molecule exhibit high exchange rates. In contrast, protons buried in the three-dimensional structure of a protein are protected from solvent exchange, and show low exchange rates. Engagement of protons in a hydrogen bond also slows down exchange by up to five orders of magnitude. *Amide proton exchange* rates are thus a measure of the accessibility of the protein backbone. To determine exchange rates of amide protons, proteins are dissolved in D_2O to start the exchange of protons with deuterons, leading to a decrease of proton resonances over time at the rate of proton/deuteron exchange. Alternatively, deuterated forms of proteins can be dissolved in H_2O, and the increasing signal from rapidly exchanging deuterons that are being replaced by protons can be observed. This approach offers the advantage of observing the resonances in the absence of other proton resonances. Amide proton exchange experiments allow the measurement of exchange rates on the time scale of seconds or slower. They have been used extensively to monitor secondary and tertiary structure formation during protein folding, and to delineate protein folding

pathways (see Section 18.5.5). Proton exchange can also be employed to probe flexibility and structure of nucleic acids, where hydrogen bond formation by base-pairing also leads to protection from exchange.

In the fast-exchange regime, when the two states interconvert rapidly ($k_{ex} \gg \omega_A - \omega_B$), the single resonance line is observed at a position and with a line width that correspond to the population-weighted average of the resonance frequencies and line widths of the two interconverting species (Figure 22.24). If resonance frequencies of the nucleus in the two states are known, the population of each state can be calculated, providing information on the equilibrium constant and hence the underlying thermodynamics. Rate constants for processes in the fast-exchange regime can be obtained from the measurement of relaxation times (Section 22.1.4, Box 22.4), which provide information on dynamics on the microsecond to nanosecond time scale. For example, transverse relaxation is enhanced in the vicinity of paramagnetic species, leading to *paramagnetic relaxation enhancement*. This effect can be used to probe distances and dynamics (Box 22.7).

BOX 22.7: PARAMAGNETIC RELAXATION ENHANCEMENT.

Paramagnetic species enhance relaxation and therefore reduce T_2 times of nearby nuclei, leading to line-broadening and eventually disappearance of their NMR resonances. Paramagnetic relaxation enhancement (PRE) can be used to locate ion binding sites in proteins. For example, Mg^{2+} binding sites can be determined by substituting with the paramagnetic Mn^{2+} ions and identifying the signals that are reduced upon binding. PRE can also be used to study iron-sulfur clusters in proteins. The introduction of spin labels (see Section 22.2) at cysteine residues can be used to probe protein-protein interfaces. Spin-labeled fatty acids and lipids enable the investigation of membrane dynamics and for the identification of protein regions that insert into or span the membrane. Due to the distance dependence of the interaction between paramagnetic species and the nucleus, PRE experiments provide distance information in the 10–20 Å range and therefore complement NOE measurements (Section 22.1.3).

22.1.8 Solid-state NMR and biology

So far, we have restricted our considerations to NMR on molecules in solution. NMR spectra from molecules in solution show narrow lines in comparison to the very broad lines in spectra of solids. Line-broadening in solid-state spectra is caused by dipolar interactions between spins and by the chemical shift anisotropy. Dipolar coupling depends on the distance r_{AB} between the two nuclei and the angle θ of the joining vector $\boldsymbol{r}_{AB}$ with respect to the magnetic field. Dipolar coupling of two spins A and B (μ_A, μ_B) leads to a shift $\Delta\nu$ in their Larmor frequency by

$$\Delta\nu \propto \frac{\mu_A \mu_B}{r_{AB}^3}\left(3\cos^2\theta - 1\right) \qquad \text{eq. 22.19}$$

Because of the strong distance dependence, dipolar coupling is strong only between nuclei that are directly bonded or immediate neighbors. In solution, the orientation of the two spins and thus the angle θ is averaged by rapid tumbling of the molecules. The average of $\cos^2\theta$ over time in this case is ⅓, and the term in parentheses in eq. 22.19 is zero. Therefore, the entire dipolar coupling term is zero and not relevant for solution NMR. In contrast, molecules in the solid state have a fixed orientation with respect to each other and to the external magnetic field, and the spin-spin interactions cannot be neglected. As a consequence, NMR lines of solid compounds are very broad. Nevertheless, solid-state NMR can provide important information on biomolecular structure and dynamics. Because of the dipolar coupling, the chemical shift and the shape of lines depend on the orientation of the

sample, which is the chemical shift anisotropy. Experiments on oriented samples and measurements at different angles of the orientation axis with respect to the external magnetic field can provide information on the orientation of helices or entire proteins within membranes. *Magic angle spinning* (MAS) experiments achieve cancellation of dipolar coupling effects by rotating the sample around its axis at a 54.7° angle with respect to the external magnetic field. The cosine squared for the magic angle is ⅓, and the time-averaged term ($3\cos^2\theta-1$) is zero. As a result, the resonances are again measured as sharp lines. High-resolution structural information on membrane proteins can thereby be obtained. Solid-state NMR experiments can also provide information on the backbone and side-chain dynamics of membrane proteins. Solid-state NMR is particularly useful to study membrane proteins in their natural environment. Solid-state methods have also been applied to fibril-forming proteins such as α-synuclein to study intrinsically disordered proteins and freeze-trapped protein folding intermediates.

22.1.9 NMR AND IMAGING

NMR can also be used as a probe for imaging purposes. In *magnetic resonance imaging* (MRI), the spatial information for each spin is encoded in its resonance frequency by employing a position-dependent external *B*-field. In a magnetic field gradient along the *x*-, *y*-, or *z*-axis, each nucleus experiences a magnetic field that now depends on its *x*-, *y*-, or *z*-coordinates. The Zeeman effect and the Larmor frequency therefore also depend on the location of the respective nucleus in this field. Thus, recording the NMR signal from each volume element of a three-dimensional specimen provides a signal whose frequency depends on the position, and whose intensity is determined by the number of nuclei it contains. By measuring the NMR signal as a function of resonance frequencies in a magnetic field gradient in the *x*-, *y*-, and *z*-directions, a map for the distribution of protons in a cell, an organ, or an entire organism can be reconstructed (Figure 22.25). Contrast can be achieved by adjusting the experimental conditions such that the measured intensity depends on the T_1- or T_2-relaxation times. For example, the relaxation times of protons in H_2O depend on the cellular environment and cover a range from 100 ms to several seconds. Freely diffusing water can thereby be clearly distinguished from water within a membrane that shows little mobility. The chemical shift can also be used as the imaging parameter.

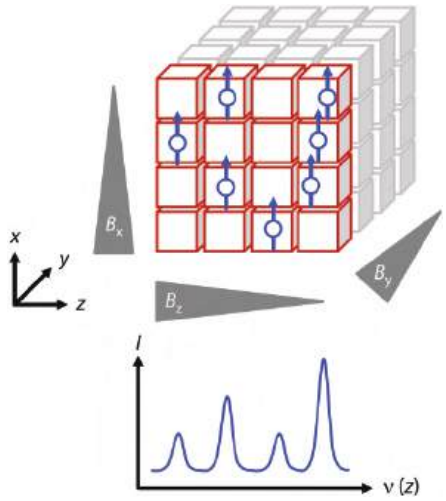


FIGURE 22.25: Magnetic resonance imaging (MRI). The three-dimensional imaging object (cube array) is placed in a magnetic field gradient. As a consequence of the field gradient in *x*-, *y*-, or *z*-direction (gray arrows), the resonance frequency of each spin (blue) depends on its coordinates along the axis of the gradient. The signal intensity at each frequency and thus at each position is proportional to the number of spins. Scanning along the *z*-axis provides signals at each frequency that are proportional to the projection of the spins in the section marked in red onto the *z*-axis. From measurements in all three dimensions, the spin density at each point of the image in three dimensions can be reconstructed.

MRI is widely used in medical diagnostics. *Functional MRI* follows changes in the brain to external stimuli and is an important tool in neuroscience and neurosurgery. Brain activity can be followed by MRI probing the oxygenation state of hemoglobin: deoxygenated hemoglobin is more paramagnetic than oxygenated hemoglobin. Deoxygenated hemoglobin therefore causes paramagnetic relaxation, and thereby alters proton resonances (paramagnetic relaxation enhancement, see Box 22.7). Active brain regions with high blood flow contain a higher level of oxygenated hemoglobin, whereas inactive regions contain more deoxygenated hemoglobin. Imaging proton resonances therefore serves as a probe for brain activity.

22.2 ELECTRON PARAMAGNETIC RESONANCE

22.2.1 Principle of electron paramagnetic resonance

Electron paramagnetic resonance (EPR) or electron spin resonance (ESR) is based on the same principles as NMR spectroscopy. Instead of nuclear spins, EPR probes the energy levels of *electron spins* and their transitions between different states. Similar to NMR-active nuclei, unpaired electrons of radicals and triplet states have a magnetic moment. Analogous to nuclear spins, any spatial orientation is possible for free electrons with spin ½ in the absence of an external magnetic field, and these states have the same energy. When an external magnetic field is present, only two orientations with respect to the field are possible, –½ and +½ (Figure 22.26). These states are of different energies. Their energy E is the scalar product of the magnetic moment of the electron and the external magnetic field B_0. The magnetic moment is the product of the g-factor of the electron g_e, the Bohr magneton μ_β, and the spin orientation m_s (½, –½) of the electron.

$$E = g_e \mu_\beta B_0 m_s \qquad \text{eq. 22.20}$$

The factor g_e is dimensionless and has a value of ~2 for a free electron. For electrons within molecules, it depends on the local magnetic field B_{local} and thus on the chemical context. The unit of the magnetic moment is provided by the Bohr magneton μ_β ($\mu_\beta = 9.27 \cdot 10^{-24}$ J T^{-1}). The resulting energy difference between the two spin states of unpaired electrons is

$$\Delta E = g_e \mu_\beta B_0 \qquad \text{eq. 22.21}$$

This difference is small compared to the energies involved in electronic transitions (Section 21.2.1), but larger than the difference between the two spin states in NMR spectroscopy (Section 22.1.1). According to eq. 22.21, transitions between the two states are induced by irradiation with electromagnetic radiation of wavelengths in the centimeter range (microwave radiation). Typical magnetic fields in EPR spectrometers are 0.1–5 T (see Table 22.2).

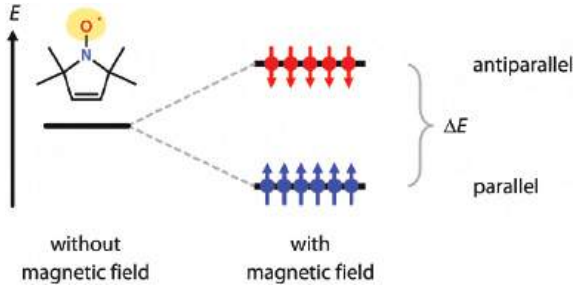


FIGURE 22.26: Energy levels and transitions in EPR. An unpaired electron, such as in a nitroxide radical, has a spin of ½. In the presence of a magnetic field, the spins align either in a parallel or in an antiparallel orientation with respect to the external field. The energies of these two spin states are different, and transitions between them can be induced by microwave radiation.

EPR spectra can be measured at a fixed, constant magnetic field while scanning the frequency of the microwave radiation. Measurements can also be done in the opposite way by keeping the microwave radiation at a constant frequency and varying the strength of the magnetic field B_0, which is the typical configuration in modern EPR spectrometers. Spectrometers are named according to the wavelength used (Table 22.2). The majority of commercially available instruments are X-band spectrometers.

TABLE 22.2:
Frequency bands of EPR spectrometers.

band	typical frequency (range) (GHz)	typical wavelength (cm)	typical *B*-field (T)
L	1.5 (1–2)	20	0.0054
S	3.0 (2–4)	10	0.11
X	9.5 (8–12)	3	0.34
Ku	17 (12–18)	1.7	0.6
Q	36 (30–50)	0.8	1.28
V	70 (50–75)	0.4	2.5
W	95 (75–110)	0.3	3.39
D	149 (110–170)	0.2	5

According to the Radio Society of Great Britain.

Due to the small energy difference between the two states, their population is almost equal, with just a tiny excess in the lower energy parallel spin orientation. The absorption signal in EPR experiments is therefore small, and EPR is rather insensitive. To achieve higher sensitivity, instruments typically measure the change in signal upon modulation of the sweeping *B*-field instead of the signal itself, and EPR spectra are recorded as the first derivative of the absorption (Figure 22.27). This derivative signal is detected at an external magnetic field B_0 where the energy difference between the spin states equals the energy of the constant microwave radiation that is applied.

Similar to nuclear spins, relaxation of excited electron spins occurs by spin-lattice or spin-spin relaxation, giving rise to the *relaxation times* T_1 and T_2, respectively. As in NMR, the T_2 relaxation time determines the line width of the EPR resonance.

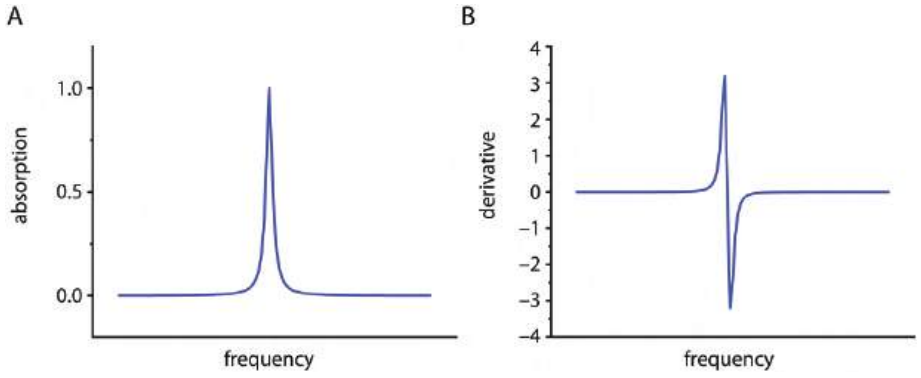


FIGURE 22.27: EPR spectrum. A: Resonance line in the absorbance spectrum. B: First derivative.

In *pulsed EPR* (Section 22.2.7), electron spins are excited by short microwave pulses, similar to FT-NMR (Section 22.1.4).

22.2.2 Spin-spin interactions: hyperfine coupling of unpaired electrons with nuclei

The *hyperfine coupling* is caused by the interaction of an electron spin S with the spin I of a neighboring nucleus. Both the electron and the nuclear spin experience the Zeeman effect in the presence of an external magnetic field. The spin-spin interaction leads to mutual effects on the local magnetic field experienced by nucleus and electron: they sense the orientation of the other spin. The important implication for EPR spectroscopy is the splitting of an EPR signal into $2I+1$ signals by a nucleus of spin I (Figure 22.28). A nucleus with $I = ½$ such as ^{1}H splits the EPR resonance into two lines, a nucleus with $I = 1$ such as ^{14}N splits it into three lines, analogous to the scalar coupling between several nuclei in NMR (Section 22.1.2 and Figure 22.4). Nitroxide spin labels that are frequently used in EPR (see Section 22.2.3) show a characteristic three-line resonance because of hyperfine coupling of the electron spin with the nuclear spin of the nitrogen. The hyperfine coupling constant is a measure of the extent of coupling and splitting, and is analogous to the coupling constant J in NMR.

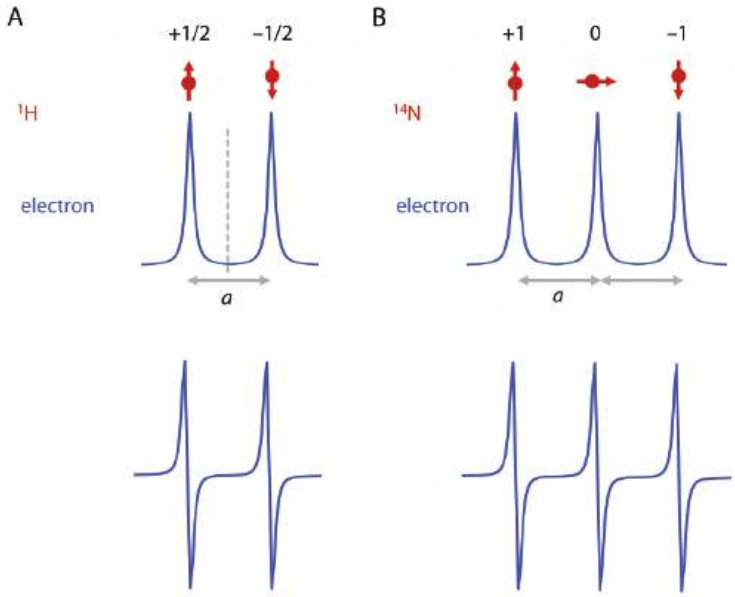


FIGURE 22.28: Hyperfine coupling. A: Splitting of an EPR resonance into two lines by an $I = ½$ nucleus such as ^{1}H. Top: absorption lines, bottom: first derivative. B: Splitting of an EPR resonance into three lines by an $I = 1$ nucleus such as ^{14}N. a is the hyperfine coupling constant.

The interaction of electron spins with nuclear spins can be exploited to determine the *hyperfine coupling constant* and to identify coupled nuclei. In *electron nuclear double resonance* (ENDOR) experiments, the electron spin is employed as a probe to detect coupled nuclei. A radiofrequency pulse induces transitions of the nuclear spins, and the effect of these transitions on the intensity of the electron spin resonance is measured. The spectrum obtained corresponds to an NMR spectrum of the coupled nuclei. By ENDOR, the hyperfine coupling constant a and the Larmor frequencies (Section 22.1.2) of the coupled nuclei can be determined. ENDOR enables the direct detection of hydrogen bonds to the radical and has been used to probe the hydrogen bond network around the active-site tyrosyl radical in ribonucleotide reductase.

22.2.3 EPR probes and spin labeling

Most biomolecules do not contain unpaired electrons and are EPR-silent. Radicals occurring in biological reactions are only transiently populated and short-lived. Exceptions are metal ion cofactors in metalloenzymes that are stable for extended periods of time and can be probed by EPR. Most of the time, however, EPR is applied to molecules that have been modified with *spin labels* or *spin*

traps. Spin labels are stable radicals, typically nitroxide radicals. A commonly used spin label is the thiol-reactive methanethiol sulfonate (MTSSL, SL for spin label), which forms disulfide bonds with Cys residues in proteins (Figure 22.29). Maleimide- and iodoacetamide derivatives of spin labels are also available. Similar to fluorescent labeling (Section 21.5.4), alternative approaches include the incorporation of spin-labeled non-natural amino acids and ligation techniques (see Section 21.5.4). For nucleic acid labeling, spin labels are introduced directly during solid-phase synthesis. Spin labels can be attached to the ribose, the phosphates or the nucleobase (Figure 22.29). A number of flexibly attached and rigid spin labels for nucleic acids have been described. For the study of membranes, spin-labeled lipids are available that contain a nitroxide moiety either attached to their hydrophobic tail or to their polar head group.

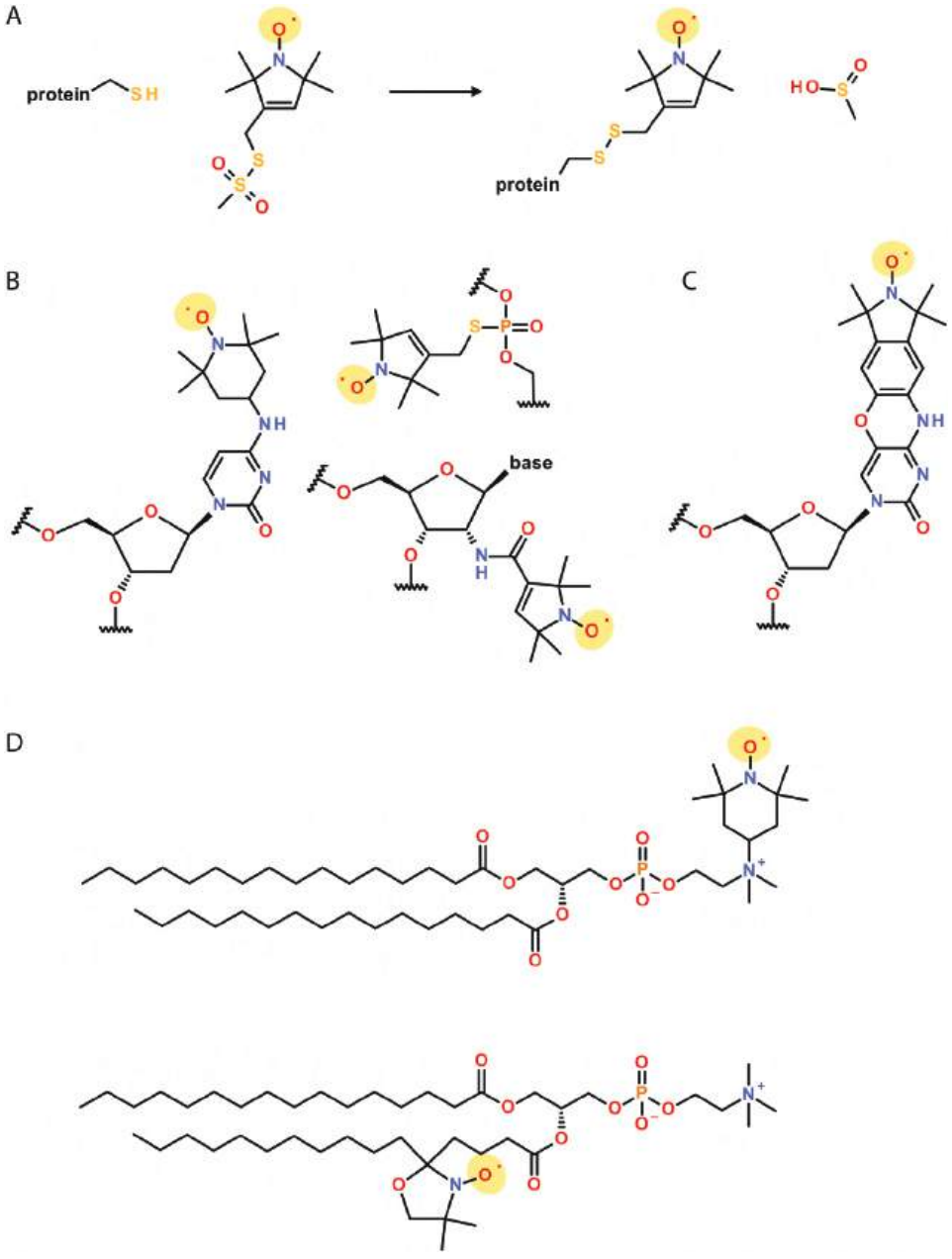


FIGURE 22.29: Spin labeling. A: Incorporation of nitroxide spin labels into proteins by disulfide exchange with methanethiolsulfonate. B: Spin-labeled nucleic acids modified at the base, the ribose and the (thio) phosphate. C: A rigid spin label for nucleic acids. D: Spin-labeling of lipids at the polar head group or within the hydrophobic tail.

Spin labels are often attached to the molecule of interest by flexible linkers. Flexible attachment provides segmental flexibility and leads to sharp lines in the EPR spectrum. Rigid spin labels show broad EPR resonances, but are useful probes for the overall mobility of the molecule of interest (Section 22.2.4) and to determine the relative orientations of the two spin centers (Section 22.2.7).

In contrast to spin labels, spin traps either convert unstable radicals into a stable radical that is EPR-active or allow recombination of an unstable radical with an EPR-active radical to an EPR-silent moiety (Figure 22.29). Spin traps are typically used to monitor radical formation. Examples of spin traps are 2,2,6,6-tetramethylpiperidine (TMP) and 2,2,6,6-tetramethyl-1-piperidinyloxy (TEMPO, Figure 22.30).

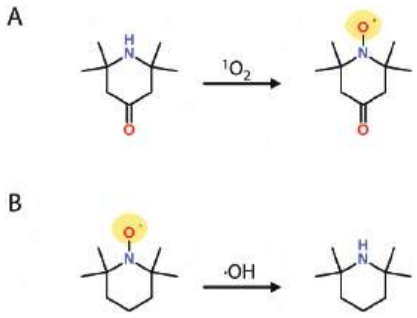


FIGURE 22.30: Spin traps. A: The EPR-silent 2,2,6,6-tetramethylpiperidin-4-one (TMP) is oxidized by singlet oxygen to the stable nitroxide 2,2,6,6-tetramethyl-1-piperidinyloxy-4-one. B: The nitroxide 2,2,6,6-tetramethyl-1-piperidinyloxy (TEMPO) reacts with hydroxide radicals to the EPR-silent 2,2,6,6-tetramethylpiperidine.

By the same principle, *i.e.* the generation or removal of a stable spin center by a transient radical, redox potentials of EPR-active cofactors can be determined in redox-titrations. The EPR signal of the cofactor is measured as a function of the buffered electrode potential of the solution during titration with oxidant or reductant, and the resulting curve is analyzed with the Nernst equation (see Section 5.4, eq. 5.15).

22.2.4 EPR AS A PROBE FOR MOBILITY AND DYNAMICS

Both *g*-factor and hyperfine coupling depend on the orientation of the spin label with respect to the external magnetic field. The extent of these orientation effects and the shape of the EPR signal depend on the mobility of the spin label. The spin label thus reports on the local mobility of the molecule of interest (Figure 22.31). For highly mobile spin labels, the orientation-dependent contributions are averaged out completely, leading to a nitroxide signal with three sharp lines. Conversely, immobile spin labels that do not move during the transverse relaxation time T_2 give rise to a broadened *rigid-limit spectrum*. In-between these extremes, intermediate shapes of the spectrum are observed.

From the shape of the EPR resonance, the rotational correlation time τ (Section 21.5.9) of the spin label can be determined as a measure of the local mobility. EPR can therefore be used to monitor local mobility and dynamics of biomolecules in solution. EPR measurements have been employed to probe membrane fluidity and dynamics, as well as protein and nucleic acid structure and dynamics. Plotting mobility data along the sequence of the protein of interest yields characteristic patterns for secondary structure elements. The dependence of the EPR signal on mobility can also be used as a probe for binding (Box 22.8) and local conformational changes. By recording EPR signals at constant magnetic field as a function of time at ambient temperature (*time-resolved EPR*), conformational changes of spin-labeled proteins can be probed at millisecond resolution. Time-resolved EPR has been used to monitor local conformational changes of spin-labeled bacteriorhodopsin in purple membranes during the photocycle in real time.

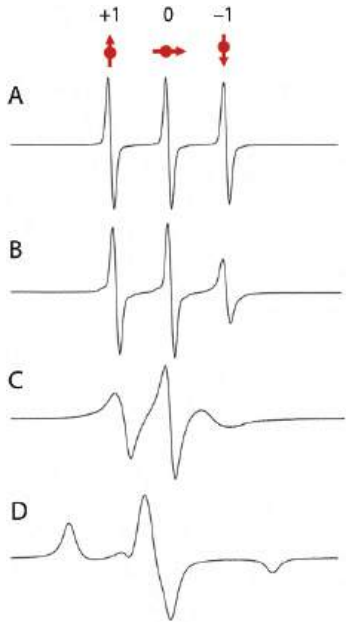


FIGURE 22.31: Effect of local mobility on EPR signals. A: EPR spectrum of the MTSSL spin label in solution. B: Spin label attached to a 15 amino acid peptide. C: Spin label attached to the same 15 amino acid peptide under conditions where it forms an α-helix. D: EPR spectrum of the frozen spin-labeled peptide. Modified from Klug, C. S. & Feix, J. B. (2008) Methods and applications of site-directed spin labeling EPR spectroscopy. Methods Cell. Biol. 84: 617–658, reprinted with permission.

BOX 22.8: EPR AS A PROBE FOR BINDING.

The shape of an EPR spectrum depends on mobility and local dynamics. Binding of ligands to spin-labeled molecules therefore often leads to a change in the EPR spectrum. Provided that the two spectra are measurably different, EPR can be used as a probe for binding in titration experiments. Before saturation is reached, the measured spectrum is a linear combination of the reference spectra of the free and bound forms. By describing the measured spectra as a weighted sum of the reference spectra, the fractions of free and bound spin-labeled molecules can be determined. Binding curves can then be analyzed to determine the K_d value (see Section 21.5.5.1). This method is also useful to determine partitioning constants of peptides and proteins into lipid bilayers.

The effect of spin label mobility on the shape of the measured EPR resonance is always dependent on the time scale of the experiment, *i.e.* on the frequency range used for excitation. The overall tumbling of spin-labeled proteins is a slow process with respect to the inverse spectral width in high frequency EPR (250 GHz) and does not contribute to the line shape. At these frequencies, more rapid motions such as backbone fluctuations of proteins become experimentally accessible and can be studied. The frequency of the microwave radiation therefore determines the time scale of motions that fall into the slow, intermediate, and rapid exchange regime (see Section 22.1.7), and that can be detected. Multi-frequency EPR takes advantage of these different time windows to study a wide range of processes occurring on different time scales.

22.2.5 EPR AS A PROBE FOR ACCESSIBILITY

EPR lines are broadened in the presence of paramagnetic compounds such as oxygen or paramagnetic ions, due to a decrease in relaxation time. Paramagnetic compounds can therefore be used to probe the accessibility of spin labels in an approach similar to fluorescence quenching (Section 21.5.7). This method has been employed to map the orientation of membrane proteins within the lipid bilayer. For such studies, oxygen and Ni^{2+}-containing compounds are typically used as paramagnetic reagents. Because of its hydrophobicity, oxygen partitions to the center of the membrane and into hydrophobic pockets of proteins, whereas its concentration in the aqueous phase is low. In contrast, Ni^{2+} compounds such as nickel (II) ethylenediamine acetate are hydrophilic and do not partition much into lipid bilayers. The effect of these reagents on the relaxation time of the spin label and the line width of its EPR signal allows distinction between sites within the membrane, in the hydrophobic core of the protein, or on the solvent-accessible surface. The intrinsic relaxation time of the spin label in the particular position is measured in a reference measurement in the presence of nitrogen.

The different partitioning of oxygen and hydrophilic paramagnetic compounds into the membrane leads to inverse concentration gradients of these reagents within the lipid bilayer, an effect that can be used to determine the immersion depth of the spin label within the lipid bilayer (Figure 22.32).

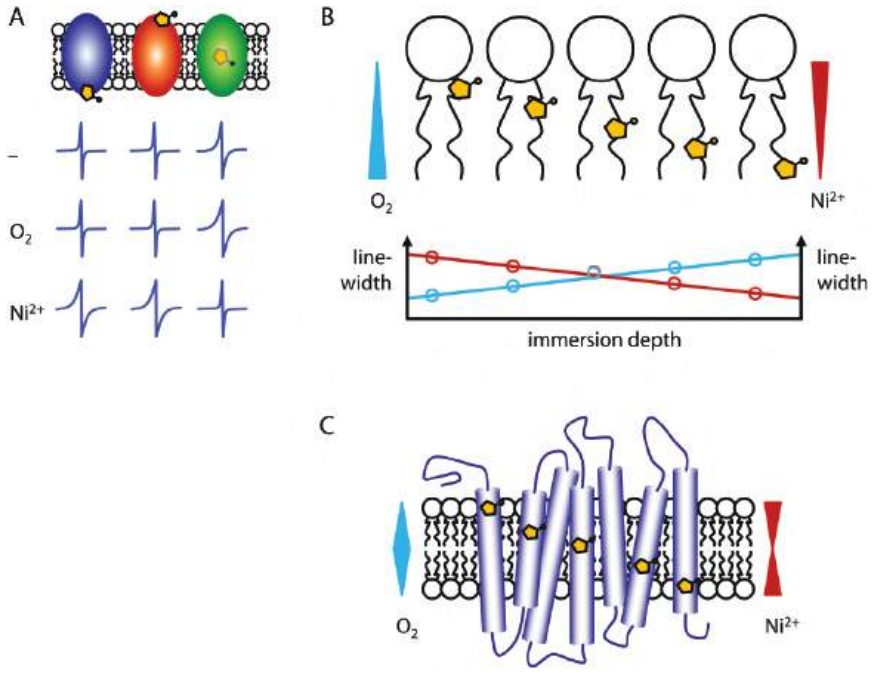


FIGURE 22.32: Relaxation enhancement by paramagnetic compounds and positioning of proteins within a membrane. A: EPR lines for a spin-labeled membrane protein with the spin label attached to the surface (left, middle) or with the spin label inserted into the lipid bilayer (green). Resonances for solvent-exposed spin labels are not affected by relaxation enhancement and line-broadening in the presence of oxygen, but line-broadening by nickel compounds is observed. Resonances for spin labels deeply immersed into the membrane are broadened due to paramagnetic relaxation enhancement in the presence of oxygen, but not in the presence of nickel compounds. Note that for simplicity only the central resonance line is shown. B: Calibration experiment for the determination of the immersion depth of spin labels within the lipid bilayer. Using membranes that contain spin-labeled lipids with known immersion depths, the response to oxygen (blue) and nickel compounds (red) can be determined as a function of the immersion depth in EPR experiments. The concentration gradients of oxygen and nickel compounds are indicated by the colored triangles. C: From the calibration experiment, responses of resonances for spin-labeled proteins to the presence of paramagnetic compounds can be converted into the immersion depth to determine the orientation and position of the protein within the lipid bilayer.

In a calibration experiment, the response of spin-labeled lipids with the spin labels attached to different positions and with known immersion depths in membranes to both paramagnetic reagents is determined. The response is either measured as a change in line width, or the collision frequencies of the spin label and the paramagnetic reagent are determined directly from relaxation times and the saturation behavior. The dependence of the response on the depth provides a calibration curve that can be used to determine the depth of a spin label attached to a membrane protein. Depths determined for different positions of the spin label then provide information on the orientation and positioning of integral membrane proteins or membrane-associated proteins.

22.2.6 Measuring spin-spin distances

The shape of the EPR signal also depends on the concentration of spins, which can be exploited to measure spin-spin distances. Inter-spin distances can be determined by EPR using three possible approaches: from the line width of cw-EPR spectra of molecules in solution, from relaxation measurements, or from oscillating patterns of magnetization detected in pulsed EPR experiments of frozen samples (Section 22.2.7).

The magnetic dipoles of two spin labels can interact through space, analogous to the dipolar interaction of nuclear spins (Section 22.1.8, eq. 22.19). The *dipolar coupling* ω_{dd} between the two electron spins depends on their distance r_{AB} and their relative orientation, described by the angle θ of the distance vector between spins A and B with the external magnetic field B_0, according to

$$\omega_{dd} \propto \frac{1}{r_{AB}^3}\left(1-3\cos^2\theta\right) \qquad \text{eq. 22.22}$$

If the two spins are close-by (below 2 nm), the dipolar coupling is large, and line-splitting can be measured by cw-EPR. For larger distances, the line-splitting caused by dipolar coupling is very small and will be hidden within the resonance line. Dipolar coupling also leads to line-broadening. The line width is related to the T_2 relaxation time, and relaxation time measurements therefore also provide information on inter-spin distance.

22.2.7 Distance determination by pulsed EPR: PELDOR/DEER

Larger inter-spin distances can be determined from pulsed EPR experiments that specifically select for dipolar coupling between electron spins. Pulsed EPR experiments to measure distances *via* dipolar coupling are called *double electron-electron resonance* (DEER) or *pulsed electron-electron double resonance* (PELDOR) experiments. The sample is excited by a series of nanosecond microwave pulses at a constant external field B_0. Such an experiment is analogous to FT-NMR with pulsed excitation. As discussed before for pulsed NMR experiments (Section 22.1.4), a 90° pulse induces transitions of spin A, and tilts the net magnetization of the spin ensemble A from the *z*-axis into the *xy*-plane. The *xy*-magnetization then decays due to the loss of phase correlation between the individual precessing spins. In *Hahn echo experiments*, the magnetization is inverted from the +*y*- to the –*y*-direction by a 180° pulse at time *t*. As a consequence, the previously diverging vectors for each precessing spin will now converge, leading to refocusing of the signal at the time 2*t* (*spin echo* or *Hahn echo*, Figure 22.33).

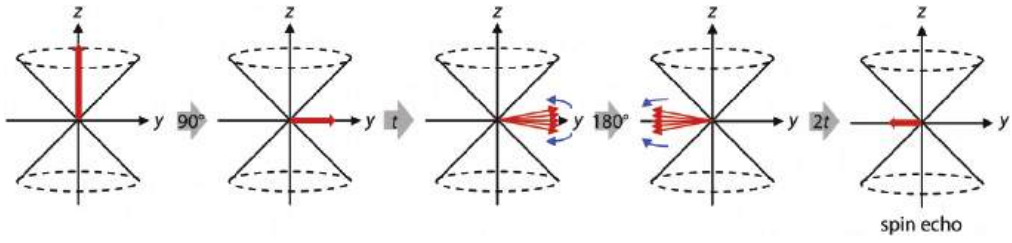


FIGURE 22.33: Hahn echo for refocusing. The net magnetization in the *z*-direction is rotated to the *y*-direction by a 90° pulse. After the time *t*, a 180° pulse inverts the magnetization to –*y*, leading to a refocusing of the magnetization vector at the time 2*t*, which generates the spin echo.

In the simplest implementation of a PELDOR experiment, a three-pulse experiment, a pump pulse at a second frequency (for spin B) will be applied at a certain time t during the experiment to invert the B spins. Because of this pump pulse and the transitions of B, the spins at the probe frequency will experience a different magnetic environment before and after the pump pulse. The size of the echo will therefore depend on the time at which the pump pulse was applied. Three-pulse PELDOR is limited by a dead time because the 90° pulse for spins A and the 180° pulse that inverts the B spins cannot be applied at the same time but have to have a minimal delay. The signal at short times t that is missed is very important to determine short inter-spin distances. Four-pulse PELDOR avoids this dead time. Here, the 180° pulse to invert spins B is applied at different times after the spin echo, and therefore temporally separated from other pulses. A subsequent additional 180° pulse at the resonance frequency of spins A then introduces a second inversion of the magnetization, and generates a refocused echo that is measured as a function of the delay time T (Figure 22.34).

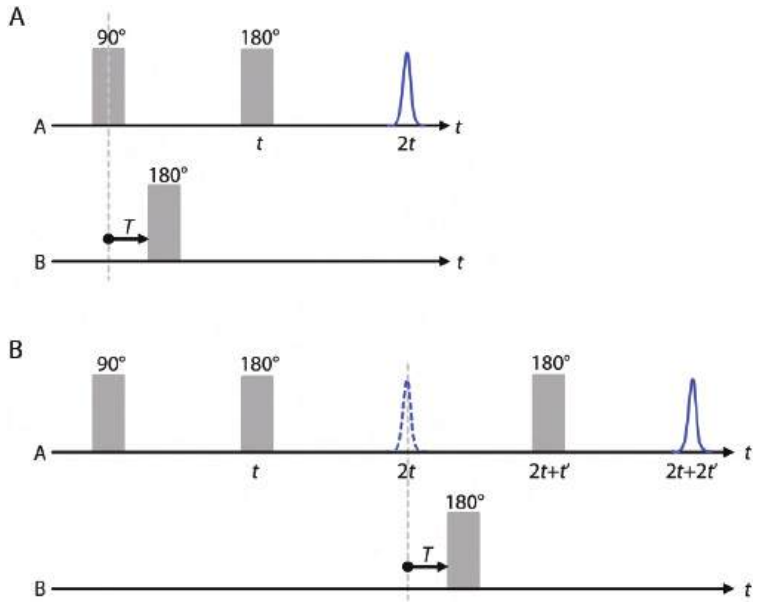


FIGURE 22.34: Three-pulse and four-pulse PELDOR. A: Three-pulse PELDOR. At the beginning of the experiments, a 90° pulse is applied at the resonance frequency of spin A. At different times T, the magnetization due to spins B is inverted by a 180° pulse. The Hahn echo of resonance A at time $2t$ (blue) is measured as a function of the time T. B: Four-pulse PELDOR. A 90° pulse is applied at the resonance frequency of spin A. A 180° pulse at time t generates a spin echo at $2t$ (dotted blue line). At different times T after this spin echo, the magnetization due to spins B is inverted by a 180° pulse. A 180° pulse at $2t + t'$ generates a refocused spin echo for resonance A at time $2t + 2t'$ (blue) that is measured as a function of the time T.

Dipolar coupling between spins A and B leads to an oscillation of the (refocused) spin-echo intensity as a function of the time when the pump pulse was applied. The characteristic coupling frequency ω_{dd} depends on the coupling strength and thus on the inter-spin distance r_{AB} (eq. 22.22). Thus, the distance between the spins can be calculated from the measured oscillation frequency of the spin echo. The measured (refocused) spin-echo intensity $V(T)$ can be described by

$$V(T) = \int_0^\infty P(r)\mathrm{d}r \int_0^1 \left(1 - p\left(1 - \cos\left(\omega_{dd}T\right)\right)\right)\mathrm{d}\theta \cos\theta \qquad \text{eq. 22.23}$$

$P(r)$ is the distribution of inter-spin distances, p denotes the fraction of spins B that are inverted by the probe pulse, and ω_{dd} is the dipolar coupling between these spins that depends on the angle θ of the distance vector between the spins with the external magnetic field B_0 (eq. 22.22). The signal $V(T)$ contains contributions from intramolecular coupling of spins on the same molecule and intermolecular coupling of spins in different molecules. For randomly distributed spins, the

contribution from intermolecular coupling can be described by a single exponential function that depends on the spin concentration. Division of the measured signal by this function yields the contribution of intramolecular dipolar spin coupling to the measured signal.

The time-dependent signal due to intramolecular spin-coupling is either Fourier-transformed into frequency space or analyzed directly. The signal is integrated over all molecules with different orientations relative to the external field B_0, and thus different values for θ. As a consequence, Fourier transformation does not yield a single line at a defined frequency, but a characteristic frequency distribution, the *Pake pattern* (Figure 22.35).

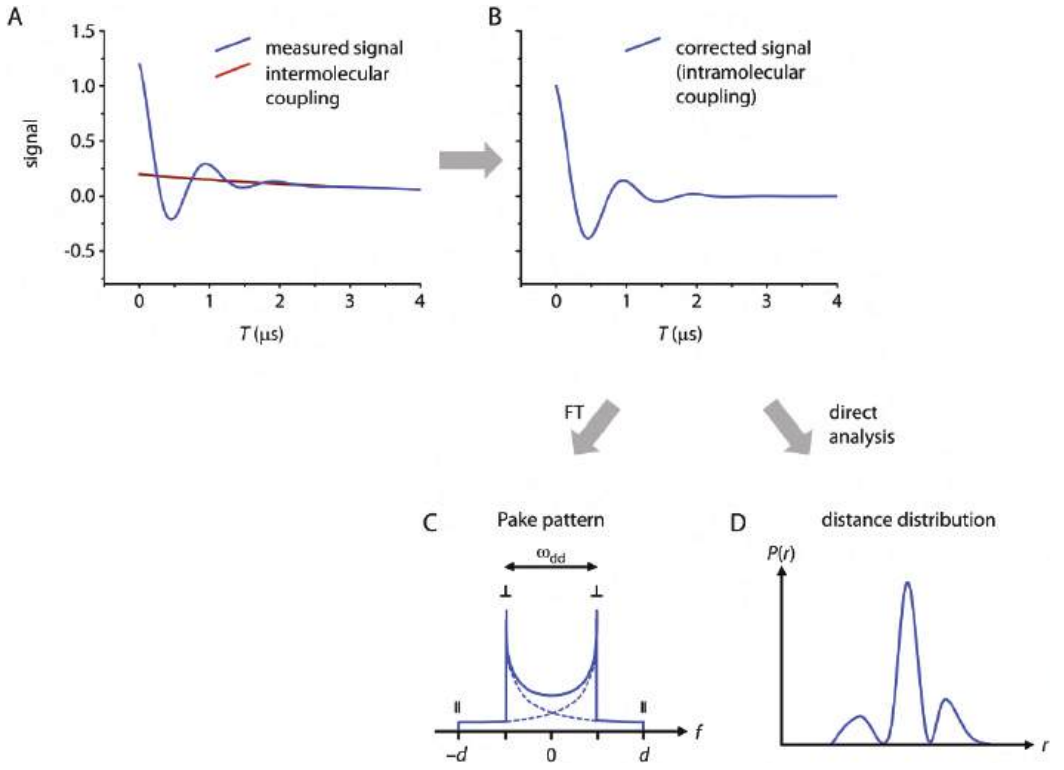


FIGURE 22.35: Distance determination from DEER/PELDOR experiments. A: Measured Hahn-echo intensity as a function of the pump pulse time *T*. B: Correction for contributions from intermolecular spin-spin interactions. The damping of the modulation depends on the distribution of inter-spin distances, which can be extracted by directly analyzing the corrected signal. C: Conversion of the signal due to intramolecular spin-spin interactions into frequency space by Fourier transformation gives a Pake pattern. The peaks correspond to spin centers with vertical and parallel orientations to the external magnetic field. The inter-spin distance is calculated from the splitting of the doublet ω_{DD}. D: Direct analysis of the corrected modulated signal by Tikhonov regularization gives a distance distribution $P(r)$.

The peaks of this pattern correspond to molecules with θ = 90°, the edges represent molecules with θ = 0°. The splitting of the doublet is ω_{dd}, from which r_{AB} can be calculated. Alternatively, the measured time-dependent signal can be directly analyzed with suitable programs that use the *Tikhonov regularization* to calculate a distance distribution. The distance distribution $P(r)$ is encoded in the damping of the modulation of the measured signal. If the spins are fixed, the relative orientation of the two spin centers, *i.e.* the angle θ between the *B*-field and the distance vector joining the spin labels, can be extracted using more involved approaches. PELDOR can also be applied to systems with more than two spin systems. Multiple distances between spin labels have been determined in PELDOR experiments of tetrameric, heptameric, and octameric membrane proteins. The modulation depth of the measured signal depends on the number of interacting spins. Determination of the number of spin centers can be used to obtain information on complex stoichiometry.

DEER/PELDOR experiments require spin label concentrations in the range of 100 μM. They are performed on frozen samples, typically at ~50–80 K. Although this precludes dynamic measurements, distance distributions determined from these experiments reflect the distribution of inter-spin distances populated in solution at the time of freezing. Time-resolved studies by pulsed EPR rely on flash-freeze methods that capture reaction intermediates or conformations at different time points. With rapid-freeze approaches relying on microfluidic mixing, reactions down to the millisecond time scale have become amenable to EPR studies.

DEER/PELDOR experiments need to measure a full oscillation of the spin-echo intensity for precise determination of distances, limiting the range of measurable distances to 1–8 nm. The need for T being long enough to cover a full oscillation requires sufficiently large relaxation times, otherwise the signal will have decayed before one oscillation is completed. The current distance record for distance determination by pulsed EPR is 10 nm (with deuterated proteins to increase the relaxation time).

Distances from EPR experiments have been used to probe conformational changes of proteins, and to validate structural models. Multiple distances, determined from molecules that are labeled in different positions, can be used to determine structures *ab initio* and to model global conformations by triangulation (Box 22.9). A recent study has demonstrated that PELDOR can also be used for distance measurements *in vivo*.

BOX 22.9: DISTANCE RESTRAINTS FROM EPR AND MODELING OF GLOBAL CONFORMATIONS.

Inter-spin distances from PELDOR measurements can be used to probe structures of large biomolecules and biomolecular complexes in solution. A combined EPR- and NMR study has used long-distance restraints from EPR and short distance restraints from NMR experiments to model the structure of the RsmZ-RsmE ribonucleoprotein complex. This 70 kDa complex consists of the 72 nucleotide RsmE RNA and three homodimers of the RsmE protein and is involved in the regulation of bacterial translation initiation. The two spin labels were introduced into the RsmZ RNA by splint ligation of two RNA oligonucleotides spin-labeled at a 4-thiouridine. The ribonucleoprotein complex was formed by adding an excess of deuterated RsmE dimers to the spin-labeled RNA. Distances of > 8 nm were determined in high power Q-band four-pulse PELDOR experiments. From the calculated distance distributions, the center of the probability distribution for each radical was determined. The distance between the two centers was used as a single restraint, and allowed to vary within a range that includes 70% probability around the maximum of the experimentally determined distance during the calculation of the structural model. The structure was modeled using both NMR short-distance restraints to define local geometry and 21 long-distance restraints from EPR to define the global conformation, starting from random coil structures. The calculated structure was validated by back-calculations of distance distributions and comparison with the experimental data. Structure calculations gave comparable results when individual restraints were omitted, which demonstrates that the structure is overdetermined by the experimental parameters. Duss *et al.* (2014) Nat. Commun. 5: 3669.

QUESTIONS

22.1 What is the magnetic field B_0 in a 600 MHz spectrometer and a 1 GHz spectrometer? The magnetogyric ratio of ^1H is $\gamma = 268{\cdot}10^6$ rad s^{-1} T^{-1}. What is the Larmor frequency of ^{13}C and ^{15}N in these spectrometers? $\gamma(^{13}C) = 67{\cdot}10^6$ rad s^{-1} T^{-1}, $\gamma(^{15}N) = -27{\cdot}10^6$ rad s^{-1} T^{-1}. What is the difference in resonance frequency for protons with a chemical shift of 2.5 ppm and 6 ppm?

22.2 What are the possible alignments for the magnetic moments of a ^{1}H and a ^{15}N nucleus with scalar heteronuclear coupling in an external magnetic field in z-direction, and what are the associated relative energies?

22.3 Explain the expected features of a 1D-NMR spectrum of ethanol.

22.4 A 1D-NMR spectrum contains four doublets. How can you tell which nuclei show scalar coupling?

22.5 Show that Γ is the full width at half maximum of an NMR line.

22.6 The dispersion of an NMR spectrum is defined as the range of chemical shifts covered by the observed resonances. Why is a large dispersion an indication for folding? What does the NMR spectrum of an intrinsically unfolded protein look like?

22.7 What NOE pattern between C_α-protons and amide protons would you expect within an α-helix and within a β-sheet?

22.8 In an NMR experiment at 10°C, a protein-ligand binding reaction is in the slow exchange regime with $k = 0.01 \cdot \Delta\omega$, and two resonance lines are observed for a proton in the ligand binding site. Can you bring the reaction into the fast exchange regime ($k = 100 \cdot \Delta\omega$) to observe these resonances as a single line?

22.9 A protein P is titrated with a ligand L, and ligand binding is observed as a change in the chemical shift of a proton in the binding site (fast exchange regime). Derive an equation for the measured chemical shift δ_{obs} as a function of total protein and ligand concentrations, P_0 and L_0.

22.10 RNA helicases consist of two domains that are connected by a flexible linker. ATP binds at the interface between the domains. In the presence of RNA, ATP binding leads to closure of the cleft between the domains, and the domains tightly interact around the bound ligands. What result would you expect in a chemical shift perturbation experiment?

22.11 You want to determine the structure of a small globular domain that is part of a 400 kDa protein. Name three possible NMR approaches and state their limitations.

22.12 Calculate the energy difference and difference in the population of the two spin states of unpaired electrons for X-band (0.34 T), Q-band (1.28 T), and D-band EPR spectrometers (5 T). $g_e = 2.00$, $\mu_\beta = 9.274 \cdot 10^{-24}$ J T^{-1}, $T = 298.15$ K. What is the wavelength of the radiation required to excite transitions? How much energy is absorbed in 500 µL of a 1 mM aqueous solution of a spin label upon saturation of the transition at each condition (in the absence of relaxation), and what is the resulting change in temperature? C_p (H_2O) = 4.18 kJ kg^{-1} K^{-1}.

22.13 Why do paramagnetic compounds decrease the relaxation times of spin labels?

REFERENCES

Alexander, N., Bortolus, M., Al-Mestarihi, A., McHaourab, H. and Meiler, J. (2008) De novo high-resolution protein structure determination from sparse spin-labeling EPR data. *Structure* **16**(2): 181–195.
· *ab initio* structure determination with distance restraints from PELDOR experiments

Altenbach, C., Flitsch, S. L., Khorana, H. G. and Hubbell, W. L. (1989) Structural studies on transmembrane proteins. 2. Spin labeling of bacteriorhodopsin mutants at unique cysteines. *Biochemistry* **28**(19): 7806–7812.
· site-specific introduction of spin labels into bacteriorhodopsin, reconstitution in lipid vesicles, and relaxation time measurements to map local dynamics

Altenbach, C., Greenhalgh, D. A., Khorana, H. G. and Hubbell, W. L. (1994) A collision gradient method to determine the immersion depth of nitroxides in lipid bilayers: Application to spin-labeled mutants of bacteriorhodopsin. *Proc Natl Acad Sci U S A* **91**(5): 1667–1671.
· collision gradient method to determine the immersion depth of spin labeled bacteriorhodopsin within the lipid bilayer by use of hydrophobic and hydrophilic paramagnetic compounds

Angulo, J. and Nieto, P. M. (2011) STD-NMR: Application to transient interactions between biomolecules-a quantitative approach. *Eur Biophys J* **40**(12): 1357–1369.
· review of theory of saturation transfer difference-NMR and quantitative applications

Argirevic, T., Riplinger, C., Stubbe, J., Neese, F. and Bennati, M. (2012) ENDOR spectroscopy and DFT calculations: Evidence for the hydrogen-bond network within α2 in the PCET of *E. coli* ribonucleotide reductase. *J Am Chem Soc* **134**(42): 17661–17670.

Baker, L. A. and Baldus, M. (2014) Characterization of membrane protein function by solid-state NMR spectroscopy. *Curr Opin Struct Biol* **27**: 48–55.
· review on applications of solid-state NMR to membrane proteins

Becker, C. F., Lausecker, K., Balog, M., Kalai, T., Hideg, K., Steinhoff, H. J. and Engelhard, M. (2005) Incorporation of spin-labelled amino acids into proteins. *Magn Reson Chem* **43**: S34–S39.
· expressed protein ligation for site-specific spin labeling

Bennati, M., Lendzian, F., Schmittel, M. and Zipse, H. (2005) Spectroscopic and theoretical approaches for studying radical reactions in class I ribonucleotide reductase. *Biol Chem* **386**(10): 1007–1022.
· review on insights into the ribonucleotide reductase mechanism from EPR studies

Bhargava, K. and Feix, J. B. (2004) Membrane binding, structure, and localization of cecropin-mellitin hybrid peptides: A site-directed spin-labeling study. *Biophys J* **86**(1 Pt 1): 329–336.
· EPR to monitor partitioning of peptides into lipid bilayers

Bode, B. E., Margraf, D., Plackmeyer, J., Durner, G., Prisner, T. F. and Schiemann, O. (2007) Counting the monomers in nanometer-sized oligomers by pulsed electron-electron double resonance. *J Am Chem Soc* **129**(21): 6736–6745.
· determination of complex stoichiometry from the presence of multiple spin centers

Borbat, P. P. and Freed, J. H. (2007) Measuring distances by pulsed dipolar ESR spectroscopy: Spin-labeled histidine kinases. *Methods Enzymol* **423**: 52–116.
· in-depth discussion of inter-spin distance determination by EPR and applications to the signaling complex of the histidine kinase CheA and CheW

Cekan, P., Smith, A. L., Barhate, N., Robinson, B. H. and Sigurdsson, S. T. (2008) Rigid spin-labeled nucleoside C: A nonperturbing EPR probe of nucleic acid conformation. *Nucleic Acids Res* **36**(18): 5946–5954.
· use of a deoxycytidine analog as a rigid spin label to probe local mobility in nucleic acids

Comellas, G. and Rienstra, C. M. (2013) Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. *Annu Rev Biophys* **42**: 515–536.
· review article on magic-angle spinning NMR and its applications to studies of amyloids

Cowburn, D. and Muir, T. W. (2001) Segmental isotopic labeling using expressed protein ligation. *Methods Enzymol* **339**: 41–54.
· review article about segmental isotope labeling for NMR using expressed protein ligation.

Derome, A. E. (1987) *Modern NMR Techniques for Chemistry Research*. Oxford, Pergamon Press.
· thorough treatment of Fourier transform NMR.

Duss, O., Michel, E., Yulikov, M., Schubert, M., Jeschke, G. and Allain, F. H. (2014) Structural basis of the non-coding RNA RsmZ acting as a protein sponge. *Nature* **509**(7502): 588–592.
· structure of the RsmZ-RsmE ribonucleoprotein complex and functional implications.

Duss, O., Yulikov, M., Jeschke, G. and Allain, F. H. (2014) EPR-aided approach for solution structure determination of large RNAs or protein-RNA complexes. *Nat Commun* **5**: 3669.
· EPR-aided structure calculation of the RsmZ-RsmE ribonucleoprotein complex

Endeward, B., Butterwick, J. A., MacKinnon, R. and Prisner, T. F. (2009) Pulsed electron-electron double-resonance determination of spin-label distances and orientations on the tetrameric potassium ion channel KcsA. *J Am Chem Soc* **131**(42): 15246–15250.
· determination of inter-spin distances and orientation of multiple spin centers in the tetrameric ion channel KscA

El Mkami, H. and Norman, D. G. (2015) EPR distance measurements in deuterated proteins. *Methods Enzymol* **564**: 125–152.
· methodology review on the determination of inter-spin distances of up to 10 nm by pulsed EPR

Felli, I. C. and Pierattelli, R. (2012) Recent progress in NMR spectroscopy: Toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* **64**(6): 473–481.
· review on applications of solid-state NMR to intrinsically disordered proteins

Fischer, M., Kloiber, K., Hausler, J., Ledolter, K., Konrat, R. and Schmid, W. (2007) Synthesis of a ^{13}C-methyl-group-labeled methionine precursor as a useful tool for simplifying protein structural analysis by NMR spectroscopy. *ChemBioChem* **8**(6): 610–612.
· synthesis of a 13C-labeled methionine precursor for isotope labeling of methionine methyl groups

Freiburger, L., Sonntag, M., Hennig, J., Li, J., Zou, P. and Sattler, M. (2015) Efficient segmental isotope labeling of multi-domain proteins using Sortase A. *J Biomol NMR* **63**(1): 1–8.
· sortase-mediated segmental isotope labeling

Gelis, I., Bonvin, A. M., Keramisanou, D., Koukaki, M., Gouridis, G., Karamanou, S., Economou, A. and Kalodimos, C. G. (2007) Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **131**(4): 756–769.
· application of methionine methyl labeling to the investigation of signal peptide binding to the SecA subunit of Sec translocase

Hagedoorn, P. L., van der Weel, L. and Hagen, W. R. (2014) EPR monitored redox titration of the cofactors of Saccharomyces cerevisiae Nar1. *JoVE* (93): e51611.
· redox titrations to determine the redox potential of EPR-active cofactors

Hagelueken, G., Ingledew, W. J., Huang, H., Petrovic-Stojanovska, B., Whitfield, C., …, Naismith, J. H. (2009) PELDOR spectroscopy distance fingerprinting of the octameric outer-membrane protein Wza from *Escherichia coli*. *Angew Chem Int Ed* **48**(16): 2904–2906.
· determination of distances between multiple-spins in the octameric membrane protein Wza

Hu, K. N. and Tycko, R. (2010) What can solid state NMR contribute to our understanding of protein folding? *Biophys Chem* **151**(1–2): 10–21.
· review on applications of solid-state NMR in protein folding and on freeze-trapped folding intermediates

Kaufmann, R., Yadid, I. and Goldfarb, D. (2013) A novel microfluidic rapid freeze-quench device for trapping reactions intermediates for high field EPR analysis. *J Magn Reson* **230**: 220–226.
· rapid-mixing device to stop reactions by freezing at different time points

Klare, J. P. (2013) Site-directed spin labeling EPR spectroscopy in protein research. *Biol Chem* **394**(10): 1281–1300.
· review article on information on protein structure and dynamics from EPR experiments

Kleckner, I. R. and Foster, M. P. (2011) An introduction to NMR-based approaches for measuring protein dynamics. *Biochim Biophys Acta* **1814**(8): 942–968.
· review article about dynamic information from NMR

Klug, C. S. and Feix, J. B. (2008) Methods and applications of site-directed spin labeling EPR spectroscopy. *Methods Cell Biol* **84**: 617–658.
· review article about spin label EPR and its applications to identify secondary structure elements, to study protein-protein- and protein-membrane interactions, folding/unfolding, and to probe protein conformation and dynamics

Krstic, I., Endeward, B., Margraf, D., Marko, A. and Prisner, T. F. (2012) Structure and dynamics of nucleic acids. *Top Curr Chem* **321**: 159–198.
· review article on applications of EPR experiments in structural and dynamic studies of nucleic acids

Krstic, I., Hansel, R., Romainczyk, O., Engels, J. W., Dotsch, V. and Prisner, T. F. (2011) Long-range distance measurements on nucleic acids in cells by pulsed EPR spectroscopy. *Angew Chem Int Ed* **50**(22): 5070–5074.

· EPR distance measurements *in vivo*

Marko, A., Margraf, D., Yu, H., Mu, Y., Stock, G. and Prisner, T. (2009) Molecular orientation studies by pulsed electron-electron double resonance experiments. *J Chem Phys* **130**(6): 064102.
· determination of the relative orientation of the two spin centers by PELDOR experiments

Marsh, D., Kurad, D. and Livshits, V. A. (2005) High-field spin label EPR of lipid membranes. *Magn Reson Chem* **43**: S20–S25.
· review article on EPR and membrane dynamics

Marušič, M., Schlagnitweit, J. and Petzold, K. (2019) RNA dynamics by NMR spectroscopy. *Chembiochem* **20**(21): 2685–2710.
· review on NMR applications to investigate the dynamics of RNA

Mollaaghababa, R., Steinhoff, H. J., Hubbell, W. L. and Khorana, H. G. (2000) Time-resolved site-directed spin-labeling studies of bacteriorhodopsin: Loop-specific conformational changes in M. *Biochemistry* **39**(5): 1120–1127.
· time-resolved EPR studies on conformational changes in bacteriorhodopsin during the photocycle

Nick, T. U., Lee, W., Kossmann, S., Neese, F., Stubbe, J. and Bennati, M. (2015) Hydrogen bond network between amino acid radical intermediates on the proton-coupled electron transfer pathway of *E. coli* α2 ribonucleotide reductase. *J Am Chem Soc* **137**(1): 289–298.
· EPR studies on the hydrogen network around the tyrosyl radical in the catalytic site of ribonucleotide reductase

Palmer, A. G., Kroenke, C. D. and Loria, J. P. (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol* **339**: 204–238.
· in-depth quantitative treatment of NMR chemical exchange to monitor conformational changes on the millisecond to second time scale

Pannier, M., Veit, S., Godt, A., Jeschke, G. and Spiess, H. W. (2000) Dead-time free measurement of dipole-dipole interactions between electron spins. *J Magn Reson* **142**(2): 331–340.
· first description of dead-time free four-pulse PELDOR and description of the Tikhonov regularization

Pervushin, K., Riek, R., Wider, G. and Wuthrich, K. (1997) Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A* **94**(23): 12366–12371.
· first description of transverse relaxation-optimized spectroscopy (TROSY) and its potential for structure determination of larger biological macromolecules

Rattle, H. (1995) *An NMR Primer for Life Scientists*. Partnership Press.
· concise introduction into the principles of NMR with a focus on structure determination of biomolecules

Rink, T., Riesle, J., Oesterhelt, D., Gerwert, K. and Steinhoff, H. J. (1997) Spin-labeling studies of the conformational changes in the vicinity of D36, D38, T46, and E161 of bacteriorhodopsin during the photocycle. *Biophys J* **73**(2): 983–993.
· relaxation time measurements of spin-labeled bacteriorhodopsin to study local protein mobility during the photocycle

Romainczyk, O., Elduque, X. and Engels, J. W. (2012) Attachment of nitroxide spin labels to nucleic acids for EPR. *Curr Protoc Nucleic Acid Chem* 7: Unit 7.17.
· direct introduction of spin labels into nucleic acids during solid-phase synthesis

Ruschak, A. M. and Kay, L. E. (2010) Methyl groups as probes of supra-molecular structure, dynamics and function. *J Biomol NMR* **46**(1): 75–87.
· review article about using methionine with labeled methyl groups for structure determination of large proteins

Schiemann, O. and Prisner, T. F. (2007) Long-range distance determinations in biomacromolecules by EPR spectroscopy. *Q Rev Biophys* **40**(1): 1–53.
· review article about the principle of EPR and applications to structure probing and distance determinations in proteins and nucleic acids

Schmidt, M. J., Fedoseev, A., Summerer, D. and Drescher, M. (2015) Genetically encoded spin labels for *in vitro* and in-cell EPR studies of native proteins. *Methods Enzymol* **563**: 483–502.
· spin labeling *via* incorporation of non-natural amino acids

Steinhoff, H. J. (2002) Methods for study of protein dynamics and protein-protein interaction in protein-ubiquitination by electron paramagnetic resonance spectroscopy. *Front Biosci* **7**: c97–110.

· review article on inter-spin distance determination by spectral moment analysis and possible applications to study protein ubiquitination

Steinhoff, H. J., Mollaaghababa, R., Altenbach, C., Hideg, K., Krebs, M., Khorana, H. G. and Hubbell, W. L. (1994) Time-resolved detection of structural changes during the photocycle of spin-labeled bacteriorhodopsin. *Science* **266**(5182): 105–107.

Tugarinov, V. and Kay, L. E. (2003) Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods. *J Am Chem Soc* **125**(45): 13868–13878.
· specific isotope labeling of isoleucine, leucine, and valine in proteins and NMR experiments for assignment of 13C and 1H resonances from these residues in large proteins

Valera, S., Ackermann, K., Pliotas, C., Huang, H., Naismith, J. H. and Bode, B. E. (2016) Accurate extraction of nanometer distances in multimers by pulse EPR. *Chem Eur J* **22**(14): 4700–4703.
· review on distance determinations for multiple spin centers in oligomers

Viegas, A., Macedo, A. L. and Cabrita, E. J. (2009) Ligand-based nuclear magnetic resonance screening techniques. *Methods Mol Biol* **572**: 81–100.
· review on NMR methods for ligand screening with a focus on practical aspects

Viegas, A., Manso, J., Corvo, M. C., Marques, M. M. and Cabrita, E. J. (2011) Binding of ibuprofen, ketorolac, and diclofenac to COX-1 and COX-2 studied by saturation transfer difference NMR. *J Med Chem* **54**(24): 8555–8562.
· saturation transfer difference experiments to characterize the interaction of ibuprofen and diclofenac with COX-I and II

Ward, R., Bowman, A., Sozudogru, E., El-Mkami, H., Owen-Hughes, T. and Norman, D. G. (2010) EPR distance measurements in deuterated proteins. *J Magn Reson* **207**(1): 164–167.
· deuteration of proteins to reduce relaxation times and to determine larger inter-spin distances

Wishart, D. S., Sykes, B. D. and Richards, F. M. (1991) Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol* **222**(2): 311–333.
· discovery of the sensitivity of chemical shifts to secondary structure

Wishart, D. S., Sykes, B. D. and Richards, F. M. (1992) The chemical shift index: A fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry* **31**(6): 1647–1651.
· assignment of secondary structure elements in proteins from Cα-proton chemical shift data

Zuiderweg, E. R. (2002) Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry* **41**(1): 1–7.
· review article about NMR approaches to investigate protein-protein interactions, including chemical shift perturbation experiments

23 Solution scattering

The techniques described in Chapters 21 and 22 are based on absorption of electromagnetic radiation by molecules. Molecules that do not absorb may interact with electromagnetic radiation and alter its propagation direction, a phenomenon termed *scattering*. Scattering depends on molecular size and shape. Depending on the size of the molecule with respect to the wavelength, different information is gained. Scattering of visible light (Section 23.1) yields molecular masses or diffusion coefficients, while scattering of X-rays and neutrons (Section 23.2) provides additional information on the shape of molecules.

23.1 LIGHT SCATTERING

Molecules that scatter light alter its propagation direction, and sometimes also its wavelength and/or phase. Light scattering can be *elastic* or *inelastic*, and *coherent* or *incoherent*. Elastic light scattering, also termed *Thomson scattering*, does not change the energy (wavelength) of the scattered light. If the scattering goes along with absorption, the energy of the scattered light is changed and we have inelastic, or *Compton scattering*. As we will see shortly, the interaction of light with matter induces a dipole that acts as a scattering center. Multiple dipoles scattering in phase are coherent scatterers, which is always the case for light that has large wavelengths compared to the size of the scattering object. If, however, the wavelength is small compared to the scattering object, the scattering dipoles may emit light of different relative phases, and incoherent scattering is observed. Thus, in most cases inelastic scattering is also incoherent. The different light scattering experiments that can be performed by adjusting the wavelength relative to the size of the object provide different information on molecules in solution, ranging from molecular mass determinations to defining their dimensions, shapes, and sometimes their internal structure.

23.1.1 Static light scattering

First, we consider the interaction of a single molecule with an electromagnetic wave. We assume that the molecule is located at the origin of a Cartesian coordinate system at (0,0,0), and exposed to linearly polarized light of a certain wavelength λ propagating along the x-axis, with the electric field vector oscillating in direction of the z-axis. The magnetic field vector oscillates perpendicular to the electric field vector in the y-direction (Figure 23.1).

We will see that the size of the particle relative to the wavelength of the incident light is of importance for the scattering properties. For now, we assume that the molecule is small compared to the wavelength of the light. The electric field vector E (Section 21.1.1) depends on the x-coordinate and the time according to

$$E(t,x) = E_0 \cos\left(2\pi\left(\nu t - \frac{x}{\lambda}\right)\right) \qquad \text{eq. 21.1}$$

E_0 is the amplitude of the electric field vector, λ is the wavelength, and ν the frequency of the incident light. At the origin, where the molecule is located, the oscillation of the electric field vector over time follows eq. 23.1:

$$E(t) = E_0 \cos(2\pi\nu t) \qquad \text{eq. 23.1}$$

An oscillating electric field $E(t)$ induces oscillations of the electrons of the molecule. The molecule therefore acts as an oscillating dipole $\mu(t)$ that depends on the polarizability α of the molecule (eq. 23.2)

$$\mu(t) = \alpha E(t) = \alpha E_0 \cos(2\pi\nu t) \qquad \text{eq. 23.2}$$

Eq. 23.2 is identical to eq. 21.9 where we have used it to describe the induced dipole in a chromophore by the electric field of the incident light. Because of its oscillating dipole moment, the molecule will function as an antenna and emit radiation. For an isotropic molecule with a centrosymmetric electron distribution, the direction of the induced dipole corresponds to the direction of the E-field vector of the incident light, in this case the z-direction. The spatial distribution of the electric field emanating from this oscillating dipole can be described as a function of the distance r from the dipole and the angle ϕ with respect to the polarization direction of the incident light:

$$E(r,\phi) = \frac{4\pi^2 \sin\phi}{r\lambda^2}\alpha E_0 \cos\left(2\pi\left(\nu t - \frac{r}{\lambda}\right)\right) \qquad \text{eq. 23.3}$$

The electric field thus depends on the sine of the angle ϕ to the z-axis. The intensity of light depends on the square of its amplitude. The intensity I_s of the scattered light is thus the square of the prefactor in eq. 23.3

$$I_s = \left(\frac{4\pi^2 \sin\phi}{r\lambda^2}\alpha E_0\right)^2 \qquad \text{eq. 23.4}$$

and the intensity of the incident light I_0 is

$$I_0 = {E_0}^2 \qquad \text{eq. 23.5}$$

From eq. 23.4 and eq. 23.5, we can calculate the ratio of the intensities of scattered and incident light as

$$\frac{I_s}{I_0} = \left(\frac{4\pi^2 \sin\phi}{r\lambda^2}\alpha\right)^2 = \frac{16\pi^4 \sin^2\phi}{r^2\lambda^4}\alpha^2 \qquad \text{eq. 23.6}$$

The relative intensity of the scattered light thus decreases with r^2, which is an inherent property of radiation from a point source. It also depends on the angle ϕ. The sine of 0° is zero, meaning there is no scattering intensity in the direction of the induced dipole ($\phi = 0°$). A large part of the scattered radiation is thus emitted in directions other than the direction of the incident light (Figure 23.1).

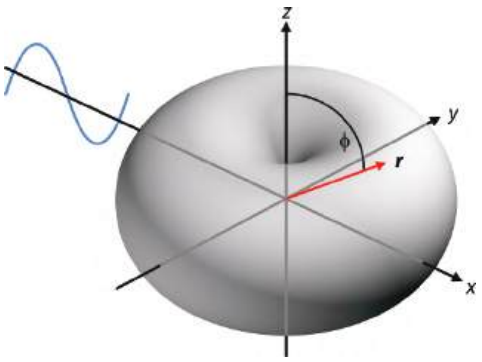


FIGURE 23.1: Scattering directions for polarized light. Incident light polarized in the *z*-direction excites oscillation of a dipole along *z*, leading to scattering into a "doughnut shape" volume. The distance from the origin to the surface equals the scattered light intensity in the direction ***r***.

We have derived the angular and distance dependence of scattered light (eq. 23.6) for polarized incident light. To generalize this equation to scattering of unpolarized light, we consider unpolarized light as a superposition of wave trains with random polarization in the *yz*-plane (Figure 23.2). Each of these individual waves elicits an electric field described by eq. 23.3, and leads to scattering intensities that are described by eq. 23.6 (Figure 23.1). The overall scattering intensity in all three directions is obtained by averaging eq. 23.3 for all directions in the *yz*-plane, which gives

$$\frac{I_s}{I_0} = \frac{8\pi^4}{r^2\lambda^4}\alpha^2\left(1+\cos^2\theta\right) \qquad \text{eq. 23.7}$$

The angle θ is the observation angle with respect to the *x*-axis. The scattering intensity according to eq. 23.7 corresponds to a dumbbell-like shape with lowest intensity in the *yz*-plane (Figure 23.2). This shape is the result of the superposition of all doughnut-like scattering surfaces that are elicited by each wave train (Figure 23.1), each rotated by a different angle around the *x*-axis because of the different polarization directions.

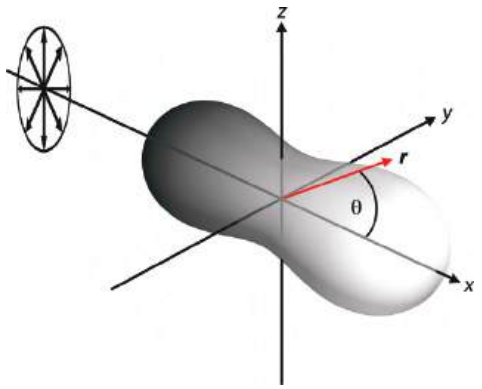


FIGURE 23.2: Scattering directions for unpolarized light. Each wave train of unpolarized light (shown as different arrows) excites the dipole in the direction of its *E*-field vector. The scattering of light in this direction is minimal. The sum of all scattered light has the shape of a dumbbell.

Eq. 23.7 describes the scattering intensity for a single molecule. In an ideal solution with *N* particles distributed over a volume *V*, each solute molecule contributes equally to the scattering intensity, with the same distance dependence and angular dependence. For ideal, infinitely diluted solutions without any interactions between the scatterers, the scattering intensity of an ensemble of molecules therefore scales with *N*:

$$\frac{I_s}{I_0} = N \cdot \frac{8\pi^4}{r^2\lambda^4} \cdot \alpha^2 \cdot \left(1+\cos^2\theta\right) \qquad \text{eq. 23.8}$$

If the scattering particles are small compared to the wavelength, this scattering process is called *Rayleigh scattering*. The *polarization factor* $1+\cos^2\theta$ in the equation for Rayleigh scattering shows that scattering is present in all directions, but it is twice as strong in the direction of the incident beam (forward and backward) as perpendicular to it. Scattering depends on the square of the molecular polarizability (α^2), which increases with molecular size. Importantly, scattering has an exquisitely strong wavelength dependence (λ^4), which is the reason for the colors of the sky we know: when viewed at large angles θ to the sun (at noon), the blue light is scattered more strongly out of the direct solar beam by particles in the atmosphere than light of longer wavelengths, reaching our eyes from all directions, so the whole hemisphere looks blue. Increased path length through the atmosphere leads to increased scattering. For this reason, the horizon appears a little whiter (less blue) than the zenith. During sunrise and sunset, the sky looks red in the direction of the sun (small θ) due to efficient removal of blue light from the direct beam of the sun by scattering.

From Rayleigh scattering experiments (eq. 23.7), the Rayleigh ratio R_θ can be measured:

$$R_\theta = \frac{I_s}{I_0} \cdot \frac{r^2}{V \cdot \left(1+\cos^2\theta\right)} = N' \cdot \frac{8\pi^4\alpha^2}{\lambda^4} \qquad \text{eq. 23.9}$$

N' is the number density of particles, which equals N/V. The Rayleigh ratio is the relative intensity of the scattered light and is related to the molecular mass of the scattering particle. In order to measure the Rayleigh ratio in practice, the polarizability α is expressed in terms of a measurable macroscopic quantity, the refractive index n, which in turn is proportional to the concentration of the scatterers and, thus, to their molecular mass M_M. The number of scatterers N in the volume V is related to the concentration c by the molecular mass M_M and Avogadro's number N_A

$$c = \frac{N \cdot M_M}{V \cdot N_A} \qquad \text{eq. 23.10}$$

The polarizability can then be expressed as

$$\alpha = \frac{n_0}{2\pi} \cdot \frac{\partial n}{\partial c} \cdot \frac{M_M}{N_A} \qquad \text{eq. 23.11}$$

n_0 is the refractive index of the pure solvent and $\partial n/\partial c$ is the *refractive index increment*, both of which can be measured. Substituting the above two equations into eq. 23.9 yields

$$R_\theta = \frac{2\pi^2 n_0^2}{N_A \cdot \lambda^4} \cdot \left(\frac{\partial n}{\partial c}\right)^2 \cdot c \cdot M_M = K \cdot c \cdot M_M \qquad \text{eq. 23.12}$$

This result shows that the Rayleigh ratio linearly depends on the molecular mass. eq. 23.12 has applications in size exclusion chromatography-multi-angle laser light scattering (SEC-MALLS), which provides inline monitoring of the molecular masses of the eluted proteins. The linear dependence of the scattering intensity on the mass of the scattering particles is also exploited in mass photometry (Box 23.1).

BOX 23.1: MASS PHOTOMETRY.

Mass photometry or interferometric scattering microscopy images single molecules according to their mass. For this type of imaging, the molecules are in a dilute solution above a refractive index interface, which is illuminated from below (Figure 23.3). Those molecules

(*Continued*)

that are close to the refractive interface scatter the incident light. The interference pattern of the low-intensity scattered light with light reflected at the interface is imaged. The scattering signal, and hence the intensity of the interference pattern, is proportional to the polarizability α of the molecules, which in turn depends on the refractive index increment $\partial n/\partial c$ and the molecular mass M_M (eq. 23.11). M_M can therefore be estimated from the intensity of the interference pattern. The accuracy of mass photometry is less than that of other methods for molecular mass determination, *e.g.* analytical ultracentrifugation (see Section 29.2.4.1). With interferometric scattering microscopy, mass determination is possible with an error of 5% and a precision of 2%. Protein masses can be determined across a wide range. The mass distribution of the sample contains information on the species present and their relative concentration.

Mass photometry is a rather simple method to analyze sample homogeneity. It can also be used to investigate protein-protein interactions, such as oligomerization or complex assembly. Proteins bound to lipids or sugars and protein-DNA complexes have also been resolved with this technique, demonstrating its applicability to biomolecular interactions in general.

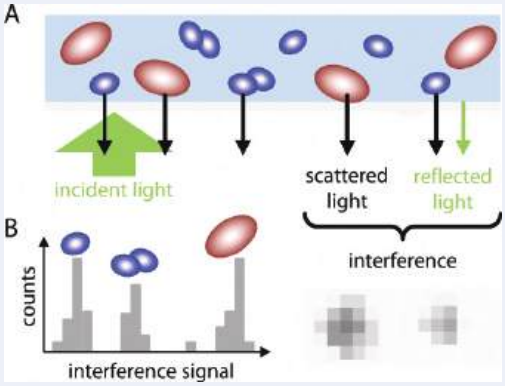


FIGURE 23.3: Principle of mass photometry. A: Single molecules (blue: monomers and dimers, red: monomer) in solution (light blue) are illuminated (green). Those molecules close to the refractive interface (gray) scatter the incident light. The interference pattern of scattered (black) and reflected (green) light is detected (bottom right). B: The intensities in the interference image are proportional to the mass of the scattering entity. The scattering intensity distribution can be converted into a mass distribution, providing information on the masses of the species present (center of the distribution) and their relative concentrations (area of the distribution).

What happens if the scattering molecules are large compared to the wavelength of the incident light? In this case, dipole oscillations are induced in several parts of the molecule that are an appreciable distance apart from each other. Each dipole will act as an antenna, and the scattered light from each dipole will come from a different origin and with a different phase. The individual scattering centers are fixed within the molecule, so their motion is correlated by the tumbling of the molecule, and hence their scattering cannot be considered independent anymore. Light that is scattered from different parts of the molecule will travel different path lengths until it reaches the detector. Therefore, the scattered light from individual scattering centers within one molecule gives an overall interference pattern for each molecule, determined by its *form factor* (see Section 23.2.4). Because of the random orientation of scattering molecules in solution, the overall observed scattering is the scattering from individual molecules averaged over all directions. The information content of the scattering experiments now extends to molecular shapes and dimensions and even the internal structure of the molecule (see Section 23.2.4.4). If the scattering is performed on a sample with ordered molecules, *i.e.* a crystal, the scattering patterns of each molecule interfere with

each other, yielding an overall scattering that is termed the *diffraction pattern*. Such patterns are no longer determined by the form factors of the individual molecules but the *structure factor* of the crystal lattice (see Chapter 24).

23.1.2 Dynamic light scattering

Instead of measuring the angular dependence of the scattering intensity, we can also follow the scattering intensity from a small volume element of our macromolecule in solution over time at a fixed angle. In this case, the scattering intensity will fluctuate because of the fluctuations in local concentration of the molecule due to Brownian motion. While on average there are N scattering molecules within the observed volume element, random thermal motion leads to fluctuations in their number, associated with an increase or decrease of the scattering intensity. The fluctuations in scattering intensity thus depend on the diffusion velocity of the molecules, which in turn is related to their molecular mass. Measuring the fluctuations of scattered light at one fixed angle θ permits the determination of the size of the molecule. If fluctuations are followed at multiple angles (multi-angle light scattering, MALS), information on the size distribution in mixtures of particles can be extracted.

The (translational) diffusion coefficient D is defined as

$$D = \frac{RT}{N_A f} = \frac{k_B T}{f} \qquad \text{eq. 23.13}$$

R is the general gas constant, k_B the Boltzmann constant, T the absolute temperature, N_A is the Avogadro constant, and f the friction coefficient. The friction coefficient f depends on the size, shape, and hydration of the molecule. For spherical molecules, f depends on the radius r as

$$f = 6\pi\eta r \qquad \text{eq. 23.14}$$

For molecules of a certain molecular mass, f is smallest when the molecule is spherical: for non-spherical molecules, the friction coefficient will be larger (discussed further for analytical ultracentrifugation; see Section 29.2). The units of the diffusion coefficient D are $cm^2\ s^{-1}$ (see Section 4.1). The diffusion coefficient can be extracted from the temporal fluctuations of the scattering intensity by calculating the autocorrelation function. The fluctuation of the scattering intensity I at time t, $\Delta I(t)$, is defined as the difference between the current intensity $I(t)$ and the average scattering intensity $\bar{I}$:

$$\Delta I(t) = I(t) - \bar{I} \qquad \text{eq. 23.15}$$

At a point later in time, at $t+\tau$, the fluctuation is

$$\Delta I(t+\tau) = \Delta I(t+\tau) - \bar{I} \qquad \text{eq. 23.16}$$

The autocorrelation function relates the fluctuation at time t with the fluctuations at different time points $t+\tau$ as

$$A(\tau) = \overline{\Delta I(t) \cdot \Delta I(t+\tau)} \qquad \text{eq. 23.17}$$

The bar indicates that the average over the time t is taken for each fixed value of τ. For small values of τ, the fluctuations at t and $t+\tau$ will be of equal sign more often than of opposite sign, and $A(\tau)$ will be positive. In contrast, for large t, the probability for $\Delta I(t)$ and $\Delta I(t+\tau)$ to be of opposite sign is very high, and $A(\tau)$ will become smaller and smaller (Figure 23.4).

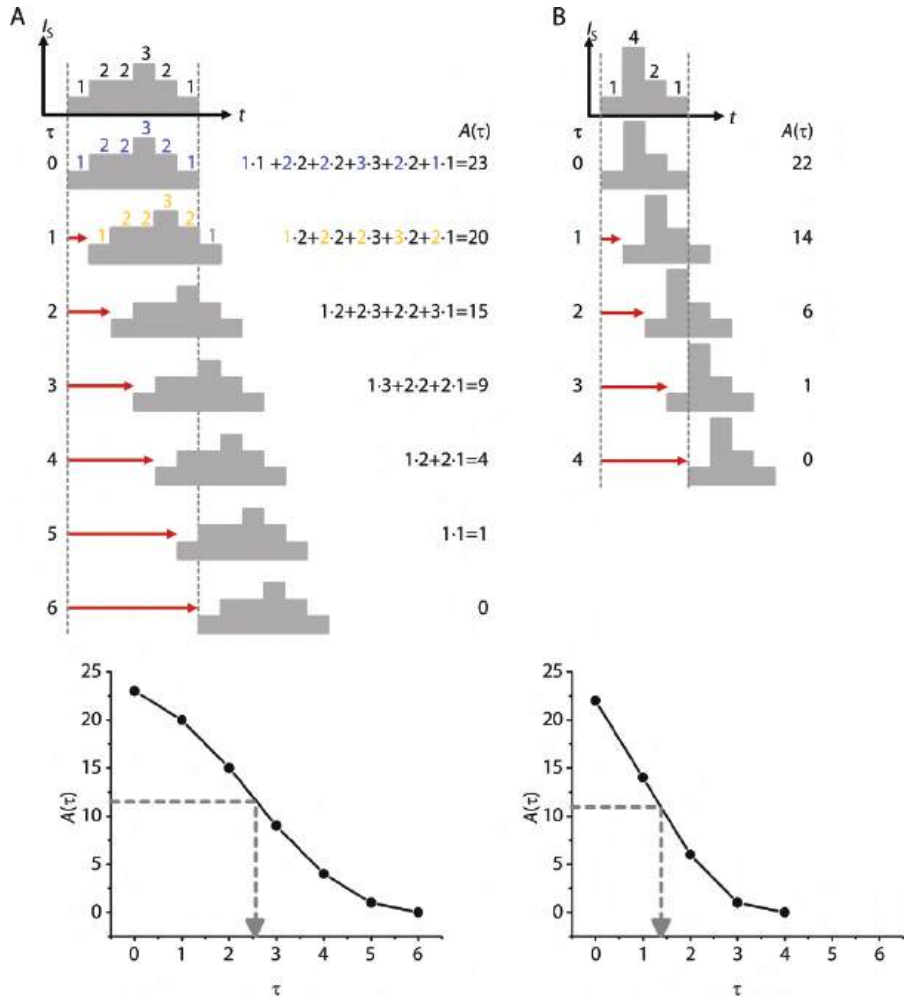


FIGURE 23.4: Fluctuations, autocorrelation function, and diffusion times. A: A molecule diffuses through the observation volume and causes a scattering signal (gray bars). After six time units, it leaves the observation volume, and the scattering signal ceases. The autocorrelation function $A(\tau)$ for this scattering pulse can be calculated by shifting the function relative to itself by 0, 1, 2, 3, 4, 5, and 6 time units, by multiplying the signal of the corresponding time points for the shifted and un-shifted function, and summing the products. The inflection point of the autocorrelation curve (bottom, 2.6 time units) corresponds to the mean time the molecule spends in the observation volume. B: A second molecule diffuses more quickly through the confocal volume and gives rise to a shorter scattering pulse. The un-normalized autocorrelation function $A(\tau)$, calculated by shifting the function relative to itself by 0, 1, 2, 3, and 4 time units, by multiplying the signal of the corresponding time points and summing the products decays to zero more rapidly. The inflection point is now at 1.4 time units, the second molecule diffuses almost twice as rapidly through the observation volume.

For a single type of scattering molecule, the autocorrelation function $A(\tau)$ at a fixed observation angle θ decays exponentially with τ (eq. 23.18):

$$A(\tau) = A_0 \exp\left(-D\frac{8\pi^2 n^2}{\lambda^2}\sin^2\frac{\theta}{2}\right) \qquad \text{eq. 23.18}$$

The diffusion coefficient D can be determined from the exponential function. Typical values for diffusion coefficients are summarized in Table 23.1. Often, the autocorrelation is plotted with a logarithmic time τ axis. The time τ at which the autocorrelation has been reduced to A_0/e corresponds to the mean diffusion time τ_{diff} of molecules through the observation volume (Figure 23.4). The diffusion coefficient D can also be calculated from τ_{diff}. We will return to the autocorrelation function in the context of fluorescence correlation spectroscopy (Section 25.6).

TABLE 23.1:
Typical diffusion coefficients *D* for biological processes.

diffusion process	*D* (cm² s⁻¹)
small molecule in aqueous solution	10^{-5}
protein in aqueous solution	10^{-6}
phospholipid in membrane	10^{-8}
protein in lipid membrane	10^{-10}

23.1.3 Raman scattering

So far, we have considered elastic scattering, in which the incident and scattered light have the same wavelength. In other words, we have assumed that no energy is exchanged when the light interacts with the molecule. In inelastic scattering, energy is exchanged, and the wavelength of the incident and the scattered light are different. This can be caused by changes in the polarizability due to internal vibrations of the molecule. For example, the overall polarizability of diatomic molecules may change with the vibration of the connecting bond. If this is the case, the polarizability α can be expressed as a function of bond vibration with the frequency ν_{bond} as

$$\alpha(\nu_{bond}) = \alpha + \beta\cos(2\pi\nu_{bond}t) \qquad \text{eq. 23.19}$$

The induced dipole then depends on $\alpha(\nu_{bond})$:

$$\mu = \alpha E_0\cos(2\pi\nu t) + \beta E_0\cos(2\pi\nu t)\cos(2\pi\nu_{bond}t) \qquad \text{eq. 23.20}$$

The first term in eq. 23.20 gives rise to elastic scattering at the same frequency ν as the incident light, and thus to Rayleigh scattering. The second term corresponds to scattering at wavelengths distinct from the wavelength of the excitation light. We can rearrange the product of the cosine terms according to eq. 23.21:

$$\cos a\cdot\cos b = \frac{1}{2}\left(\cos(a+b)+\cos(a-b)\right) \qquad \text{eq. 23.21}$$

and obtain

$$\mu = \alpha E_0\cos(2\pi\nu t) + \frac{\beta E_0}{2}\left(\cos\left(2\pi t(\nu+\nu_{bond})\right)+\cos\left(2\pi t(\nu-\nu_{bond})\right)\right) \qquad \text{eq. 23.22}$$

Eq. 23.22 thus reveals that, in addition to elastic Rayleigh scattering (first term), scattering occurs at higher and lower frequencies $\nu+\nu_{bond}$ (second term) and $\nu-\nu_{bond}$ (third term; Figure 23.5). This phenomenon is called *Raman scattering*. The Raman bands at lower frequency than the excitation light occur because a vibration is excited in a molecule, and are called Stokes bands. The bands at higher frequency than the excitation light occur when a molecule in a higher vibronic state is irradiated and returns to a lower vibrational state. They are called *anti*-Stokes bands. The energy of the vibration can thus be calculated from the spectral difference between Raman bands and excitation light. To measure a Raman spectrum, the sample is illuminated with monochromatic, plane-polarized light. The scattered light is measured at a 90° angle.

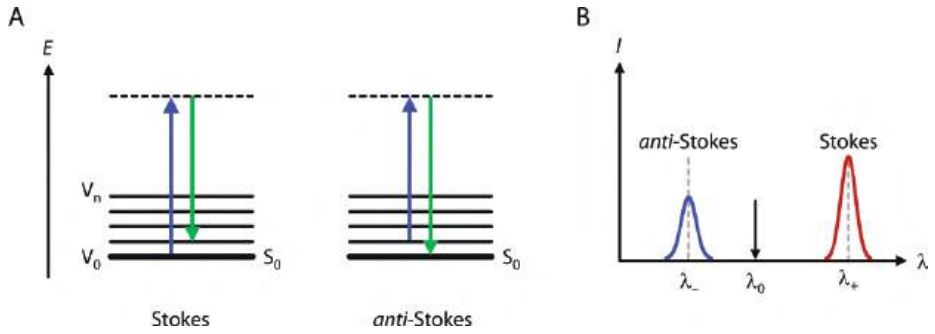


FIGURE 23.5: Raman scattering. A: Raman spectroscopy is based on inelastic scattering, caused by a change in polarizability of molecules with bond vibrations. The scattered light is of lower energy/lower frequency when a bond vibration is excited (left), giving rise to the Stokes band. If a vibronic state relaxes, the scattered light is of higher energy / higher frequency (right), giving rise to the *anti*-Stokes band. B: *anti*-Stokes and Stokes bands at λ_- and λ_+ after irradiation at λ_0.

The coefficient β in eq. eq. 23.22 is typically less than 10^{-4} of α, and Raman bands are of much lower intensity than the Rayleigh band. Nevertheless, Raman spectroscopy can be very valuable for the study of biomolecules. Raman bands are observed when the polarizability of the molecule changes with the vibration. Vibrations can also give rise to absorption of infrared light (Section 21.4). Infrared absorption occurs when the vibration leads to a change in charge distribution, *i.e.* requires a non-zero transition dipole. Because of these different requirements, Raman spectroscopy is often complementary to infrared spectroscopy. It should be noted that Raman bands from water vibrations distort fluorescence spectra and need to be corrected for (see Section 21.5.6). On the scale of Raman bands from biomolecules, however, Raman bands from water are small, and water absorption does not interfere with Raman studies of biomolecules in aqueous solution. In Raman microscopy, Raman spectroscopy is used as a probe in imaging.

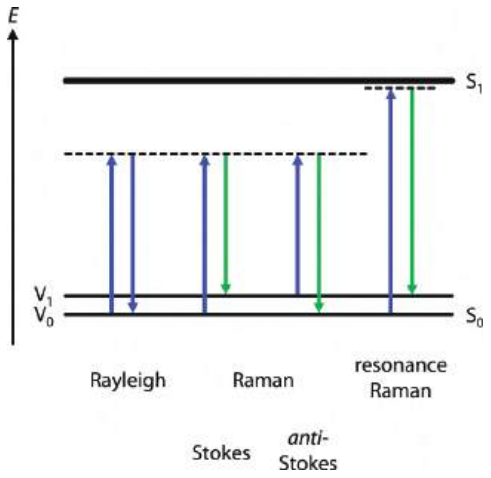


FIGURE 23.6: Rayleigh, Raman, and resonance Raman scattering. Rayleigh scattering is elastic, and the incident and scattered light are of the same wavelength. Raman scattering is inelastic, and the scattered light is either of higher (Stokes) or of lower wavelength (*anti*-Stokes) than the incident light. The energy difference excites a higher vibrational state (Stokes) or stems from the relaxation of a higher vibrational state (*anti*-Stokes). In resonance Raman scattering, the wavelength of the incident light lies within an absorption band. Here, the transition for the Stokes band is shown.

Raman spectroscopy encounters two major problems that limit its application to biomolecules. First, the intensity of Raman scattering is typically very low. Second, Raman spectroscopy reports on vibrational transitions, and the number of bands in a Raman spectrum increases with the number of

degrees of freedom of the molecule ($3N-6$ for a nonlinear molecule of N atoms, see Section 21.4.2). Such complex Raman spectra provide a signature for molecules, and can be used, for example, as fingerprints of pharmaceutical compounds. G-quadruplex structures (see Section 19.5.2.2) show different Raman spectra depending on whether they are parallel or antiparallel.

The intensity of Raman bands can be enhanced by resonance and surface effects. Resonance Raman spectroscopy exploits the enhancement of Raman signals near electronic transitions, *i.e.* near an absorption band (Figure 23.6). This enhancement is caused by the large change of polarizability associated with absorption. In resonance Raman spectroscopy, vibrations of a chromophore can therefore be observed not only with high sensitivity, but also selectively within a larger molecule or molecular assembly. Resonance Raman spectroscopy has been used to study heme or flavin groups, carotenes, iron-sulfur clusters, and the retinal chromophore in rhodopsin.

Raman scattering is also intensified near surfaces of noble metals such as gold, an effect that is exploited in surface-enhanced Raman spectroscopy (SERS). When SERS is combined with the resonance Raman effect (surface enhanced resonance Raman spectroscopy, SERRS), single-molecule sensitivity is reached. SERS fingerprinting of molecules, cells, and tissues is used in analytical applications. SERS has been used to study membrane proteins such as bacteriorhodopsin, rhodopsin, and photosynthetic complexes in model membranes or their natural membrane environment.

23.2 SMALL ANGLE SCATTERING

In Section 23.1.1 we discussed how light scattering from molecules that are small compared to the wavelength of the incident light is different from scattering by large molecules. X-rays and particle beams have wavelengths on the order of bond lengths, much smaller than the average molecule diameter. In principle, they thus can provide higher-resolution information than light scattering. In the following, we discuss the scattering of X-rays and neutrons by molecules. The intensity of scattered radiation decreases strongly at larger angles with respect to the incident light. Scattering intensity can therefore only be measured at small angles, hence the collective term small angle scattering (SAS).

23.2.1 Scattering of X-rays and neutrons

X-rays are scattered by electrons, and neutrons are scattered by nuclei, but the physical principles of the two phenomena are the same. Both scattering processes can be elastic or inelastic. We will focus on the elastic scattering of X-rays, *i.e.* no energy is transferred from the incident light onto the sample. The incident radiation induces oscillations in electrons, which then emit secondary X-rays in all directions.

In the introduction to light scattering, we derived the equation for the overall scattering intensity as a function of scattering angle and direction (eq. 23.7). The polarizability α of electrons is a function of natural constants (electron charge e, electron mass m_e, and vacuum permittivity ε_0) and the frequency ω of the electron:

$$\alpha = \frac{e^2}{4\pi \cdot \varepsilon_0 \cdot m_e \cdot \omega^2} \qquad \text{eq. 23.23}$$

The frequency ω can be expressed in terms of the speed of light c and the wavelength λ as $\omega = 2\pi c/\lambda$, and we can eliminate the polarization and wavelength from eq. 23.7 to arrive at the *Thomson equation*. The Thomson equation describes the intensity of the light scattered by a single electron I_s relative to the absolute intensity of the incident beam I_0.

$$\frac{I_s}{I_0} = \frac{e^4}{16\pi^2 \cdot \varepsilon_0{}^2 \cdot m_e{}^2 \cdot c^4} \cdot \frac{\left(1+\cos^2\theta\right)}{2r^2} = C_{Th} \cdot \frac{\left(1+\cos^2\theta\right)}{2r^2} \qquad \text{eq. 23.24}$$

The constant factor in the Thomson equation, $C_{Th} = 7.94 \cdot 10^{-30}$ m^2, is a very small number, reflecting the minute scattering intensity from a single electron. However, the sum of all electrons present in macromolecular solutions and crystals will scatter light at detectable intensities. The unit m^2 of C_{Th} indicates a *cross-section*, a measure of the likelihood of a collision event between particles (Section 29.1.5.5) or photons. The constant is proportional to the scattered light intensity and is related to the extinction coefficient in absorption. Its square root amounts to 2.82 fm, the *Thomson scattering length* b_e. b_e signifies the amplitude of the scattered radiation at $\theta = 0$ per unit solid angle (in steradian). In other words, it is an absolute measure for the scattering power of a single electron.

$$b_e = \frac{e^2}{4\pi \cdot \varepsilon_0 \cdot m_e \cdot c^2} = \frac{\mu_0 e^2}{4\pi m_e} = 2.82 \text{ fm} \qquad \text{eq. 23.25}$$

In eq. 23.25, μ_0 is the vacuum permeability, which depends on the vacuum permittivity and the speed of light as $1/(\varepsilon_0 c^2)$. b_e is also known as the *classical electron radius*. Although electrons are typically considered as point charges with no spatial extent, b_e can be viewed as the radius of a sphere of charge e whose potential energy equals the potential energy of a particle with the electron rest mass m_e. The scattering length b_{atom} of an atom with atomic number Z normally scales with the number of electrons:

$$b_{atom} = Z \cdot b_e \qquad \text{eq. 23.26}$$

Table 23.2 lists the X-ray and neutron scattering lengths for a few biologically interesting isotopes. While the linearity of b_{atom} for X-rays is evident, neutron scattering lengths do not scale with Z. Neutron scattering lengths describe the more complicated interaction between a neutron beam and atomic nuclei. For example, the magnitudes of the neutron scattering lengths of hydrogen (^{1}H) and deuterium (D) are comparable to those of much heavier elements, such as sulfur and uranium (Table 23.2). Hydrogen atoms are therefore detectable by neutron scattering. Note also the negative value for the ^{1}H neutron scattering length compared to D. This sign difference allows discrimination between ^{1}H and D by neutron scattering. We will come back to this point in the description of small angle neutron scattering (SANS; Section 23.2.5).

TABLE 23.2:
Coherent scattering lengths (in fm) for X-rays and neutrons.

		scattering length (fm)	
isotope	***Z***	**X-rays**	**neutrons**
^{1}H	1	2.9	–3.7
^{2}H (D)	1	2.9	6.7
^{3}H (T)	1	2.9	4.8
^{12}C	6	17.1	6.7
^{13}C	6	17.1	6.2
^{14}N	7	20.0	9.4
^{15}N	7	20.0	6.4
^{16}O	8	22.8	5.8
^{31}P	15	42.8	5.1
^{32}S	16	45.6	2.8
^{238}U	92	262	8.4

The relative scattering power of an atom is expressed in terms of the ratio of the scattering lengths of the atom and the electron (for X-rays; eq. 23.27). This ratio is the *atomic scattering factor f*, also known as the *atomic form factor*. In other words, *f* is the ratio of the amplitudes of the scattered waves of the atom and a single electron:

$$f = \frac{\text{amplitude of wave scattered by the atom}}{\text{amplitude of wave scattered by a single electron}} = \frac{b_{atom}}{b_e} \qquad \text{eq. 23.27}$$

Mathematically, the atomic scattering factor is the Fourier transform (Section 31.15) of the electron density of a spherical atom. It is a function of the scattering angle θ. At θ = 0 and in the absence of absorption of the incident X-rays or other side effects, f_0 is equal to the number of electrons in the scattering particle (Z for atoms). The magnitude of f decreases with larger scattering angles following a function of the form $\exp(-\sin^2\theta)$ because at larger scattering angles the waves emitted by each of the Z electrons in an atom will have larger phase shifts relative to each other, and thus show increasingly destructive interference (Figure 23.7). Because of this strong angular dependence, scattering intensity can only be measured at small angles (small angle scattering, SAS).

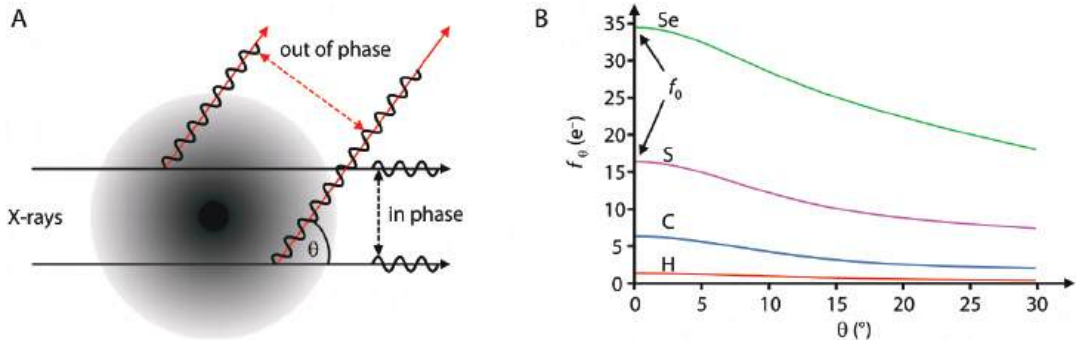


FIGURE 23.7: Atomic scattering factors. A: Elastic scattering from two regions of the electron cloud of an atom introduces phase shifts depending on the scattering angle. In the forward direction (black, θ = 0), there is no phase shift because the path difference between the diffracted beams is zero. The path difference, and hence the phase shift, increases with increasing θ. B: Resulting angle dependence of the atomic scattering factor for a few elements. At θ = 0 the atomic scattering factor equals the number of electrons (H: 1, C: 6, S: 16, Se: 34).

The scattering length $b_{molecule}$ of a molecule consisting of N atoms can be calculated as the sum of all atomic scattering lengths:

$$b_{molecule} = \sum_{i=1}^{N} b_{atom,i} \qquad \text{eq. 23.28}$$

In analogy to the atomic form factor, this *molecular form factor* describes the overall relative scattering amplitude of a particle (molecule) in ideal solution relative to an electron. Rays scattered by different atoms in the same molecule will interfere with each other (Figure 23.8) and provide information on the extent of the molecule (its form). Eq. 23.28 only holds for ideal solutions, where the scattered radiation interferes only within the same molecule. In non-ideal solutions the molecules interact with each other, and scattered light from different molecules shows additional interference. This scattering is not described by a form factor any more, but by the *structure factor*. The structure factor is particularly important for ordered molecular arrays such as crystals, where it describes the expected intensity distribution of light diffracted by the array. Because scattering of solutions and diffraction by crystals are based on the same physical process, they share many aspects, which we will point out throughout this Chapter and in Chapter 24 on X-ray crystallography.

For the diluted and practically ideal solutions used in small angle scattering (SAS) studies, the scattering length per volume is an informative parameter. The volume of a molecule can be calculated from the density ρ, the molar mass M, and the Avogadro constant N_A. Division of the scattering length of a molecule by its volume yields its *scattering length density* SLD, sometimes also referred to as *scattering density* ρ.

$$SLD = \frac{b_{\text{molecule}}}{V_{\text{molecule}}} = \frac{\rho \cdot N_A \cdot b_{\text{molecule}}}{M} \qquad \text{eq. 23.29}$$

Scattering densities are useful to judge the amount of expected scattering from a particular sample. Carbohydrates, lipids, water, proteins, and nucleic acids all have characteristic X-ray and neutron scattering densities (Table 23.3). Nucleic acids have the highest scattering densities and thus give stronger signals for SAS. The scattering densities for X-rays and neutrons in Table 23.3 follow similar trends as the atomic form factors given in Table 23.2: substances with higher electron density scatter X-rays more strongly, and the neutron scattering densities for the hydrogen isotopes have the opposite sign.

TABLE 23.3:
Scattering densities and electron densities for biological substances.

	X-rays		neutrons	
substance	**SLD (10^{-4} nm^{-2})**	**ρ_e (e nm^{-3})**	**SLD (10^{-4} nm^{-2})**	**% D_2O**
H_2O	9.4	334	–0.56	0
D_2O	9.4	334	6.4	100
proteins	11.5–12.6	410–450	0.02–0.03	40–45
DNA	16.5	590	0.04	65
RNA	16.7	600	0.05	72
carbohydrates	13.8	490	0.03	47
lipids	8.7–9.6	310–340	0.01–0.04	10–14

Data from Perkins, S. J. (1988). *Modern Physical Methods in Biochemistry, Part B* (A. Neuberger & L. L. M. van Deenen, Eds.), with permission. For neutron scattering it is more convenient not to state the nuclear density but the percentage of D_2O that is needed to experimentally match the density of the given substance (the *neutron match point*).

The differences in X-ray and neutron scattering densities between water (H_2O and D_2O) and biological substances indicate that both are essentially difference measurements: the scattering *contrast* between solvent and solute is measured (Figure 23.8). The contrast depends on the difference in the scattering densities of the macromolecule and the solvent (Table 23.3). Given all other parameters are identical, subtracting the X-ray scattering data of a buffer control from the scattering data of a sample leaves the scattering of the macromolecule or, more precisely, of the macromolecule plus its sphere of hydration (see radius of hydration; Box 23.2). For SANS, the change in nuclear density between solvent and solute is negligible in water (Table 23.3), but changes dramatically in the presence of D_2O. A technique called *density matching* or *contrast variation* is used in SANS to determine the shape of macromolecules (see Section 23.2.5.2).

23.2.2 SAS INTENSITY DISTRIBUTION

Atomic scattering factors decrease strongly as a function of scattering angle due to increased negative interference of the beams scattered at different parts of the atom (Figure 23.7). The same is true for scattering by molecules. The scattering into the same direction from two different points of a molecule will result in two beams of different path length and, hence, different phase. These beams interfere with each other and the detector only registers the *interference pattern* or *scattering pattern* (Figure 23.8).

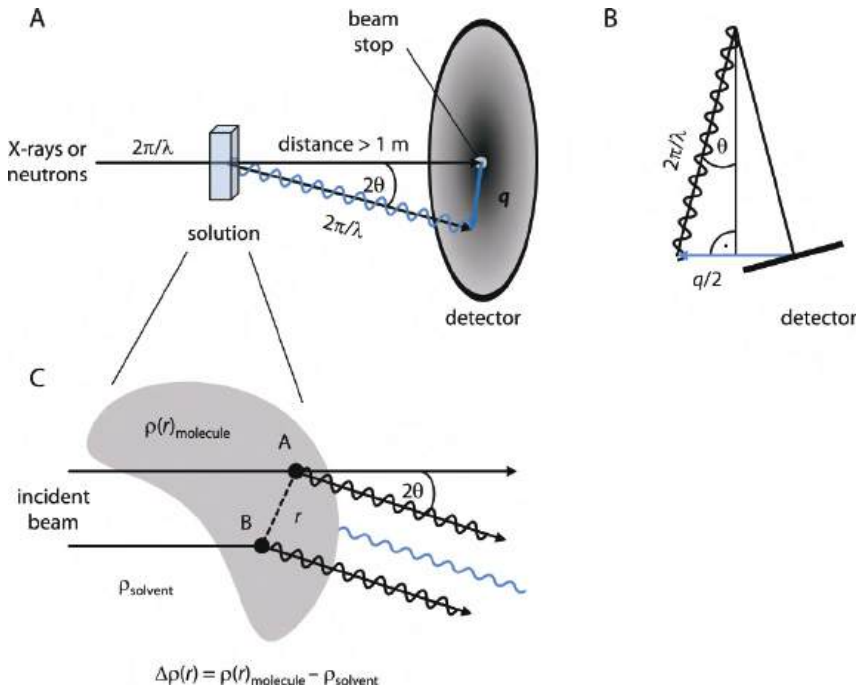


FIGURE 23.8: Schematic SAS experiment. A: The incident (direct) beam of wavenumber $2\pi/\lambda$ illuminates a sample in solution. The detector is at several meters distance from the sample and is protected from the intense direct beam by a beam stop. The angle between the direct and the scattered beams is defined as 2θ. The scattering pattern on the detector is radially symmetric because the molecules in solution adopt all possible orientations. Darker shading toward the beam stop represents higher scattering intensity. B: The scattering vector $\boldsymbol{q}$ is related to the scattering angle and the wavenumber $2\pi/\lambda$ by trigonometry. C: Scattering from two different points A and B in a molecule (gray shape) leads to negative interference for $\theta > 0$, indicated by the blue sine wave of lower amplitude.

Within molecules, both the scattering angle θ and the distance r between the scattering points influence the path difference. For θ approaching zero, the path difference is also zero, and hence interference is absent with no loss of scattering (intensity I_0) in the forward direction. At angles $\theta > 0$, the scattered intensity is conveniently described by the *scattering vector* $\boldsymbol{q}$ (Figure 23.8), which is the result of the interference of all beams scattered into certain directions by all points of the molecule.

The incident beam can be represented by the *spatial frequency* $2\pi/\lambda$, which represents the number of complete waves (2π) per unit distance. Together with the scattering angle θ, the spatial frequency defines the direction and amplitude of the scattering vector $\boldsymbol{q}$. The magnitude q of the scattering vector is related to the scattering angle θ by trigonometry (Figure 23.8).

$$q = \frac{4\pi}{\lambda}\sin\theta \qquad \text{eq. 23.30}$$

The parameter q, also known as the *momentum transfer*, is both a measure of resolution and of the change in directional momentum of the scattered photons: the larger the change in directional momentum, the higher resolution information can be gathered from the experiment (Figure 23.9). The unit of q is inverse length ($Å^{-1}$ or nm^{-1}). We will see later that q is related to the Bragg resolution d in crystallography (see eq. 24.7) by $q = 2\pi/d$. The inverse length of the momentum transfer highlights that scattering happens in *reciprocal space*, in contrast to the *real space* of molecules. The mathematical connection between the two spaces is the Fourier transformation (see Section 31.15).

Because the molecules in solution have random orientations, $I(q)$ is proportional to the scattering of a single particle averaged over all of these orientations. The scattering from molecules in random orientations produces an isotropic pattern, which means it has radial symmetry about the

incident X-ray beam (Figure 23.8). Different molecular shapes lead to different $I(q)$ distributions (Figure 23.9).

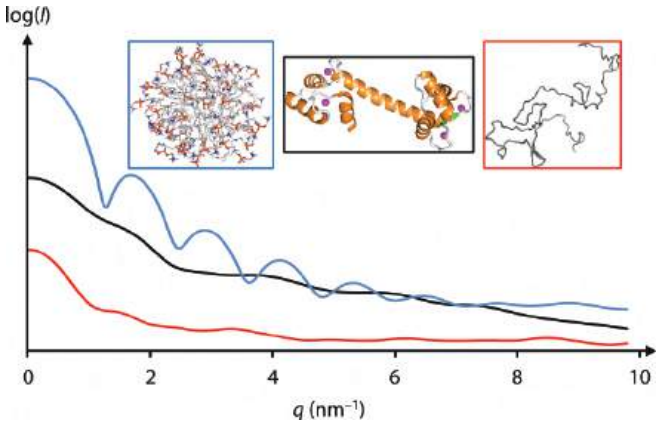


FIGURE 23.9: Examples of scattering curves from molecules with different shapes. Scattering curves for a small spherical micelle (blue), the elongated, helical Ca^{2+}-binding protein calmodulin (black), and an extended random coil (red). The more isometric the molecule, the more features appear in the intensity distribution at high q values. Since the scattering intensity decreases strongly with increasing q values, log(I) is plotted. Calculated from coordinates using the program CRYSOL at www.embl-hamburg.de/biosaxs/atsas-online.

For a monodisperse and ideal solution, *i.e.* a highly diluted solution without interactions between freely diffusing solute molecules, the scattering intensity as a function of small q values is given by the *Guinier approximation* (Figure 23.10).

$$I(q) = I_0 \cdot e^{-\frac{R_g^2 q^2}{3}} \qquad \text{eq. 23.31}$$

R_g is the *radius of gyration* and I_0 is the total scattering intensity at $\theta = 0$. Mathematically, R_g is the quadratic mean of the distances r_i between atoms and the center of mass, weighted by the contrast m_i of the atoms:

$$R_g = \sqrt{\frac{\sum m_i r_i^2}{\sum m_i}} \qquad \text{eq. 23.32}$$

R_g is a useful measure for the degree of particle size and anisometry, *i.e.* its degree of asymmetry. For example, spheres of radius R follow the linear dependence $R_g = \sqrt{0.4} \cdot R$, while cylindrical rods of radius R and length L (nucleic acids) follow $R_g^2 R^2/4 + L^2/12$. R_g is closely related to another quantity, the *radius of hydration* R_h. Methods such as SAS are sensitive to the layer of water bound to the surface of macromolecules and measure R_h rather than R_g (Box 23.2).

BOX 23.2: R_g *VERSUS* R_h.

Dissolved macromolecules bind a layer of water on their surface. This sphere of hydration corresponds to ca. 0.3 g H_2O per gram protein and is about 15% more densely packed than bulk solvent (see Figure 18.41). The added solvent changes R_g to R_h, the *hydrodynamic radius* or *radius of hydration*. R_h equals the radius of an equivalent hard sphere that diffuses with the same velocity as the molecule under observation (including its hydration sphere). R_h can be larger or smaller than R_g, depending on the shape and flexibility of the

(*Continued*)

particle. For spherical particles, the hydration shell increases R_g slightly. In theory, a single layer of water molecules on the surface of a spherical molecule will increase its radius by ca. 0.34 nm, but in practice several layers are present, and the two radii are then roughly related by $R_h \approx 1.3 \cdot R_g$. R_h is measured by methods that are sensitive to diffusion, such as light scattering, analytical ultracentrifugation (see Section 29.2), gel permeation chromatography, electrophoresis, and also SAXS.

The Guinier approximation is valid only for q values satisfying the relation $R_g \cdot q < 1.3$. Thus, the larger the molecule, the larger R_g, and the smaller the q values where $I(q)$ follows the Guinier approximation. Eq. 23.31 provides a convenient method to calculate R_g from the measured scattering curves. Taking the logarithm of eq. 23.31 yields the *Guinier equation*.

$$\ln\left(I(q)\right) = I_0 - \frac{R_g^2}{3} \cdot q^2 \qquad \text{eq. 23.33}$$

The plot of $\ln(I(q))$ as a function of q^2 is known as a *Guinier plot* (Figure 23.10). A fit of the Guinier equation to the linearized scattering data at small q values yields R_g of the particle from the slope.

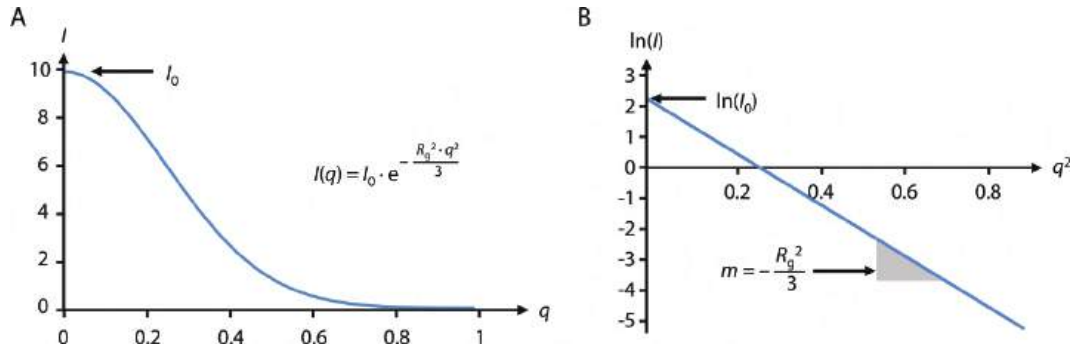


FIGURE 23.10: Theoretical solution scattering curve. A: Exponential intensity decay as a function of scattering angle. An arbitrary I_0 of 10 and R_g of 5 nm were used for this calculation. B: Linearization of the intensity in a Guinier plot allows retrieval of R_g and I_0 by a linear fit. In practice, the plot will only be linear for q values satisfying $R_g \cdot q < 1.3$.

The second parameter obtained from the Guinier plot is I_0, the total forward scattering intensity as predicted from Rayleigh scattering (Section 23.1). I_0 is an important parameter in SAXS because (after buffer subtraction) it is proportional to the square of the effective number of electrons in the sample volume that was illuminated by the incident beam. Thus, I_0 is also proportional to the mass concentration of the solutes, which in principle allows computing molecular masses (Section 23.2.4.2), independent of the particle shape. Unfortunately, I_0 cannot be measured directly because this forward-scattered beam lies in the direction of the incident beam. The beam stop not only blocks the direct beam, but also inevitably truncates the measured scattering curves at small θ-angles (Figure 23.8). Therefore, the lowest resolution data have to be measured carefully to obtain an accurate value of I_0 at the intercept $\theta = 0$ ($q = 0$) from the linear fit.

23.2.3 Distance distribution function

Since the Guinier approximation to calculate R_g and I_0 is limited to small q values where only few data are available, it is preferable to include all of the scattering data in their calculation. The *pair distance distribution function* $P(r)$, or short *distribution function*, provides such an approach. $P(r)$ describes the electron distribution of a particle in real space, and the corresponding intensity distribution $I(q)$ in reciprocal space is given by a Fourier transform of $P(r)$:

$$I(q) = 4\pi \cdot \int_0^{D_{max}} P(r) \cdot \frac{\sin(q \cdot r)}{q \cdot r} dr \qquad \text{eq. 23.34}$$

D_{max} is the maximum distance r between two electrons in the molecule. Since $I(q)$ is measured, $P(r)$ can be calculated from the measured $I(q)$ data by a numerical *Fourier inversion*, also known as the indirect transformation procedure of Glatter. This operation converts the discrete number of measurement points of the scattering curve into the continuous $P(r)$ function.

$$P(r) = \frac{r}{2\pi^2} \cdot \int_0^{\infty} q \cdot I(q) \cdot \sin(q \cdot r) dq \qquad \text{eq. 23.35}$$

The beauty of the $I(q)$ and $P(r)$-Fourier pair is that both I_0 and R_g can now be calculated using all of the available measurement data according to

$$I_0 = 4\pi \cdot \int_0^{D_{max}} P(r) dr \qquad \text{eq. 23.36}$$

and

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 \cdot P(r) \cdot dr}{\int_0^{D_{max}} P(r) \cdot dr} \qquad \text{eq. 23.37}$$

Ideally, the I_0 and R_g values calculated from the distribution function should coincide with the values from the Guinier plot: their comparison provides a useful consistency check for the data. In any case, I_0 and R_g values calculated from $P(r)$ are more reliable since they use all available SAXS/SANS data.

The different scattering curves for molecules with different shapes (Figure 23.9) are caused by different distance distributions between the electrons in these molecules. $P(r)$ is an autocorrelation function of the electron density or, in other words, it represents an account of how often which distance between two electrons is present in the molecule. It now becomes clear why the point $P(r) = 0$ represents the maximum particle dimension D_{max}: it is impossible for two electrons within a molecule to have a distance exceeding D_{max} (Figure 23.11).

The shape of the $P(r)$ function contains information on the shape of the molecule. Globular macromolecules have unimodal distance distributions, while multi-subunit proteins may display several distance distribution maxima. The first maximum corresponds to a set of short distances within a domain while other maxima represent sets of distances between the domains. Since there are, on average, fewer large distances within a molecule than shorter distances, $P(r)$ tails off towards D_{max}. Flexible proteins with many extended conformations have larger tails in their $P(r)$ functions, corresponding to a higher number of large distances.

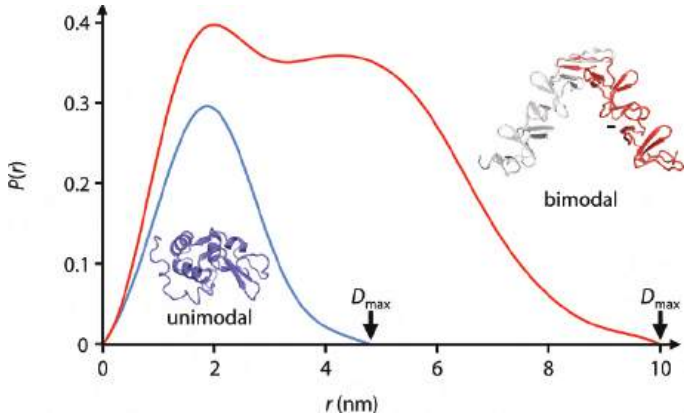


FIGURE 23.11: *P*(*r*) electron pair distribution functions. Globular, single-domain proteins such as lysozyme have unimodal distance distributions (blue curve). The dimeric lytic amidase (LytA; red curve) shows bimodal distance distribution. For the dimeric LytA, the large maximum at shorter distances represents the distances within one monomer. The set of larger distances between the monomers (shown as red and gray ribbons) is the second maximum on the right. Data are from the sasbdb.org database (accession IDs SASDA96 and SASDBJ4). D_{max} is the maximum distance between electrons in the particle.

$P(r)$ is not only useful for comparing different proteins, but also for detecting conformational changes within a molecule. These include large structural changes upon ligand binding, protein and nucleic acid folding, or oligomerization. For example, a change in relative domain positions in an oligomer affects many distance pairs between the domains, drastically changing the shape of the $P(r)$ distribution.

23.2.4 Small angle X-ray scattering

Small angle X-ray scattering (SAXS) and X-ray diffraction experiments use the same X-ray sources. The generation of X-rays is described in Section 24.4.1. SAXS is one of the more powerful methods to assess molecular flexibility, without the size limitations of NMR and EM.

23.2.4.1 SAXS experiment

Unless the aim is to analyze a mixture of macromolecules, for instance a monomer/oligomer equilibrium, SAXS samples should be > 90% homogeneous. SAXS measurements are optimally performed at several concentrations, which have to be known accurately with < 10% error because they are used to normalize the scattering data. If the scattering curves collected at different concentrations superimpose after the scaling by concentration, the molecule does not change its aggregation state over the concentration range, and the sample behaves like an ideal solution. A typical sample for SAXS is about 15 µL in a quartz capillary at a concentration of 1–10 mg mL^{-1} for proteins. Nucleic acid samples can be less concentrated because of their higher electron density. To maximize the difference in electron density between solvent and solute, the contrast, a buffer of low ionic strength (< 0.5 M) is used with as few additives (< 0.5 mM) or co-solvents (< 5% glycerol) as possible. X-ray-induced radiation damage of the sample can be a problem in SAXS and X-ray crystallography. Radiation damage is assessed by repeated measurement of scattering from the same sample and comparison of the scattering curves, which should be identical. Co-solvents such as glycerol, formate, and thiols act as radical scavengers and can reduce radiation damage. If radiation damage poses a problem despite the presence of radical scavengers, a flow cell may be used that continuously offers fresh material to the X-ray beam at the expense of more sample. In case detergents have to be used to stabilize proteins for SAXS measurements, their concentration needs to be below the *cmc* (Section 18.4.3) to avoid contributions to X-ray scattering from micelles.

An exposure at a modern synchrotron beamline lasts about 1 s, depending on the wavelength, the beam size, and the molecular mass and concentration of the macromolecule. Scattering intensities are corrected for buffer contributions by subtracting the buffer scattering measured in a separate

experiment. This allows measurement of SAXS data in the range of 0.1 $nm^{-1} < q < 6\ nm^{-1}$, equivalent to scattering intensities over three orders of magnitude. Small q values are required for accurate R_g and D_{max} determinations, while larger q values (higher resolution data) are needed for reliable structure modeling (Section 23.2.4.4).

Two of the first parameters that are accessible from SAXS data are I_0 and R_g from a Guinier plot (Figure 23.10). An accurate estimate for R_g is obtained by limiting the data analysis to the *Guinier region* ($R_g \cdot q < 1.3$). Since R_g is not known initially, several rounds of fitting may be required to identify the best data region for R_g estimation. For aggregated samples, the Guinier region is not linear, so R_g and I_0 will be different compared to their values obtained from the $P(r)$ function. In this case, different buffer conditions or purification methods (filtering, centrifugation) are required to establish a monodisperse sample. Other parameters that can be calculated from the scattering data are the excluded volume, the molecular mass M_M, and a Kratky plot.

23.2.4.2 Excluded volume and molecular mass

The excluded volume V of a molecule is the volume occupied by the molecule plus its sphere of hydration that cannot be penetrated by bulk solvent (Box 23.2). The volumes of large (> 70 kDa) globular macromolecules can be approximately calculated from the low-resolution scattering data using the *Porod equation* (eq. 23.38).

$$V = 2 \cdot \pi^2 \cdot \frac{I_0}{\int_0^\infty q^2 \cdot I(q) \cdot \mathrm{d}q} \cdot \qquad \text{eq. 23.38}$$

The integral in the denominator of eq. 23.38 is the *Porod invariant*. Since the integration is over all q values, the Porod invariant is proportional to the total scattering intensity I_0, and the volume calculation does not require data normalization by the solute concentration. V is also obtained from *ab initio* shape modeling (Section 23.2.4.4), and both methods should yield consistent values. Based on the fairly uniform density of macromolecules, the excluded volume in nm^3 is converted into a molecular mass in kDa using the rule-of-thumb relation $M_M \approx 0.6 \cdot V$. For non-spherical, asymmetric, and smaller macromolecules, the Porod volume calculation is inaccurate. In this case, the molecular mass is determined from the total forward scattering intensity I_0 of a solution with known concentration c:

$$\frac{I_0}{c} = \frac{N_A \cdot M_M}{\mu^2} \cdot \left(1 - \frac{\rho_0}{\rho_{sample}}\right)^2 \cdot \qquad \text{eq. 23.39}$$

The parameter μ is the ratio of the molecular mass to the number of electrons of the molecule, typically 1.87 for proteins. ρ_0 and ρ_{sample} are the average electron densities of the solvent and the sample (see Table 23.3). If the sample is not monodisperse, the molecular mass calculated with eq. 23.39 will be a mass-average of all the constituents of the sample:

$$\bar{M}_M = \frac{\sum_i N_i \cdot M_i^2}{\sum_i N_i \cdot M_i} \qquad \text{eq. 23.40}$$

where N_i is the number of molecules of molecular mass M_i.

23.2.4.3 Kratky plot

The Kratky plot visualizes the compactness of a macromolecule and can be directly calculated from the scattering curve (Figure 23.12). It is simply a different way to plot the intensity data as a function of the magnitude q of the scattering vector **q**, namely as the product $q^2 \cdot I(q)$ *versus* q. The Kratky plot for a globular, folded, and compact protein is bell-shaped. Unfolded and extended molecules, on the other hand, show a plateau in the Kratky plot that does not decay toward the baseline at high q values.

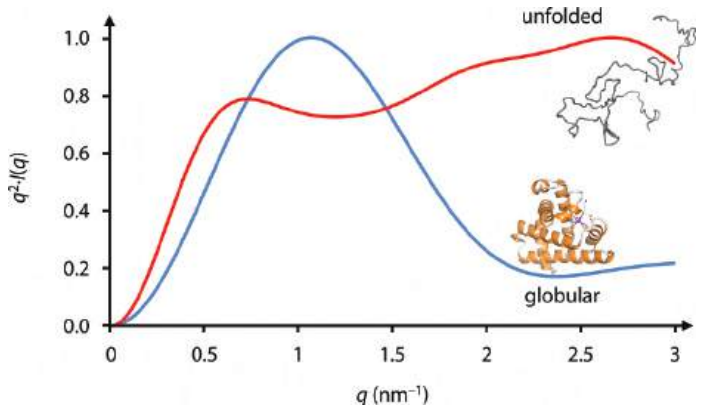


FIGURE 23.12: Kratky plot. $q^2 \cdot I(q)$ is plotted as a function of q. Globular proteins (blue curve) exhibit a peak in this plot, while unfolded (extended) proteins give rise to a plateau.

By way of the Kratky plot, SAXS provides similar information on protein folding and unfolding as far-UV CD spectroscopy (Section 21.3).

23.2.4.4 Modeling of scattering curves

The modeling of the entire scattering curve by a molecular envelope, termed *shape reconstruction*, extends the scattering analysis beyond the mere extraction of molecular parameters. Due to the limited data-to-parameter ratio inherent in SAXS data, care must be taken not to over-interpret the data. Often, modeling of molecular shapes from scattering curves uses crystal or NMR structures of macromolecules as rigid bodies. Flexible macromolecules pose a special problem for SAXS modeling because multiple molecular envelopes can be found that fit the data equally well and therefore undermine the reliability of the shape reconstruction. If experimental data from other methods are available, hybrid approaches (Box 20.1) can improve the shape reconstruction.

If no structural information is available, *ab initio* modeling of low-resolution particle shapes can be performed without other information, directly from the scattering pattern. The starting model is either a random coil polypeptide or an array of several thousand beads that are much smaller than the resolution of the SAXS data, and, in total, approximate the macromolecule as a *dummy atom model*. If the resolution of the SAXS data is < 2 nm, the number of beads may be increased to match the number of amino acids or nucleotides (see Section 20.1.4). By using coarse-grained modeling and simulated annealing procedures (Sections 20.1.3 and 20.1.4), a compact arrangement of the beads that best fits the measured data is searched for. The quality of the fit can be assessed by the *normalized discrepancy function*, which is a measure of the differences between the measured data I_{obs} and the calculated scattering data I_{calc} from the final structural model:

$$\chi^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[\frac{I_{obs}(q_i) - c \cdot I_{calc}(q_i)}{\sigma(q_i)} \right]^2 \qquad \text{eq. 23.41}$$

N is the number of data points I_{obs} with their associated errors σ_i, and c is a scaling constant between the measured and calculated intensities. Minimizing χ^2 during the molecular dynamics simulation explicitly takes the errors of the data into account. The experimental errors have to be correctly estimated during data processing. As a result of the fitting, a set of different models is obtained that all describe the scattering data equally well. The smaller the differences between the individual models, the more reliable is the fitting result. SAXS shape reconstructions are compared using the *normalized spatial discrepancy* (NSD). NSD calculations can be done in the absence of coordinates, only based on object shape. For each point in structure S1, all distances to all points in structure S2 are calculated, and the minimum distances are noted. The calculation is repeated for all points in S2, then all distances are added and normalized by the average distances between neighboring points in the structures:

$$NSD_{S1,S2} = \sqrt{\frac{1}{2}\left[\frac{1}{N_1 \cdot \overline{d_2}^2}\sum_{i=1}^{N_1}\delta_{S1i,S2}{}^2 + \frac{1}{N_2 \cdot \overline{d_1}^2}\sum_{i=1}^{N_2}\delta_{S2i,S1}{}^2\right]} \qquad \text{eq. 23.42}$$

Identical structures would have NSD = 0. A low NSD value allows calculation of an average structure and to extract the most probable structure from the ensemble. For NSD > 1, the structures diverge in a systematic manner, indicating that there might be structures of very different shapes in the sample. For instance, a dynamic monomer-oligomer system will scatter as the sum of monomer and oligomer, weighted by the mole fractions. SAXS data for pure monomer at low concentration and oligomer at high concentration can be collected, and the mole fractions of monomer and oligomer at intermediate concentrations can be extracted by a weighted fit using the references. The shape reconstruction will then give envelopes for both the monomer and the oligomer. For flexible molecules that adopt a plethora of conformations in solution, the shape reconstruction is more difficult. Sometimes, the calculated shapes can be grouped into clusters that may represent sub-populations of conformers in solution. The molecular shapes calculated from SAXS data contain an ambiguity. Both enantiomorphs (mirror images) of the shape envelope will describe the measured data equally well, and additional information from other techniques is required to decide on which enantiomorph is correct.

23.2.5 Small angle neutron scattering

SAXS and small angle neutron scattering (SANS) are in many aspects complementary techniques, and all SAXS data evaluation methods also work for SANS data. In addition, SANS offers the method of *contrast variation*, which is based on the opposite sign of the ^{1}H and D neutron scattering lengths. Another advantage of SANS compared to SAXS is the virtual absence of radiation damage because neutrons generate much less radicals than X-rays do, which enables long measurement times. However, while SAXS beam time is widely accessible at synchrotrons, SANS experiments are limited to a few large neutron production facilities.

23.2.5.1 Generation of neutrons

Neutrons, like any elementary particles, can be viewed both as particles or waves. The wave-particle dualism of quantum mechanics posits that a *de Broglie wavelength* can be assigned to a neutron of mass m_n traveling at velocity v:

$$\lambda = \frac{h}{m_n \cdot v} \qquad \text{eq. 23.43}$$

where h is the Planck constant. For SANS measurements, slow or *thermal neutrons* are required, which travel with an average velocity of a few km s^{-1}. These neutrons are not easily available – they are mainly produced in two types of large facilities known as *fission reactors* and *spallation sources*. In a fission reactor, neutrons are generated in a controlled chain reaction where thermal neutrons hit a suitable fuel, for instance enriched ^{235}U:

$$^{235}U + n \longrightarrow {}^{236}U \longrightarrow {}^{89}Kr + {}^{144}Ba + 3\,n \qquad \text{scheme 23.1}$$

The fission of a uranium nucleus yields fast neutrons, which are decelerated by a moderator, often H_2O or D_2O. The neutrons leave the reactor at a continuous flux of up to $8{\cdot}10^{14}$ neutrons per cm^2 and second.

The other major method to produce neutrons is spallation. A suitable target, *e.g.* lead (Pb) or Bismuth (Bi), is bombarded by protons that were accelerated to relativistic speeds using a synchrotron. The impact on the target atom produces 20–30 neutrons per absorbed proton. The Swiss SINQ source produces a flux of 10^{14} neutrons per cm^2 and second.

The average energy of neutrons from a water moderator at ambient temperature is about 25 meV. Most experiments require a monochromatic neutron beam, which is generated either by velocity selection (time of flight) or by Bragg reflection (diffraction) of the neutrons (see eq. 24.7) from a

single crystal. For velocity selection, a beam cutter rotating at a certain speed selects neutrons for both a specific velocity and direction. The result is a parallel, or *collimated*, beam of a specified wavelength. In Bragg reflection, a monochromator passes only neutrons of a specific wavelength on to the experiment.

23.2.5.2 Contrast variation

The information gained from a neutron scattering experiment of a macromolecule in solution depends on the difference density between the solute and the solvent. In SANS the nuclear difference density can be adjusted by a method termed *contrast variation* or *solvent matching*. We have introduced the neutron scattering length as a measure of the strength of the interaction between a neutron and a nucleus (Section 23.2.1). Recall that the scattering lengths of ^{1}H and D differ not only in their magnitude but also in their sign (Table 23.2). ^{1}H has a large negative scattering length, which means that the neutron beam scattered from ^{1}H is phase-shifted by 180° relative to beams scattered by any other atom type in the molecule. As usual, the total scattered intensity is the result of interference of all beams scattered by all atoms. The phase shift by ^{1}H therefore has a major effect on the total scattering. By contrast, deuterium has a large positive neutron scattering length, and therefore its contribution to the total scattering is almost opposite to that of ^{1}H. If, for example, we prepare a deuterated macromolecule and dissolve it in H_2O, the scattering of the solvent will be very different from the scattering of the solute, greatly increasing the contrast between the two. The reverse experiment, dissolving a solute containing ^{1}H in D_2O, is also possible. By adjusting the H_2O/D_2O ratio to match the solvent scattering density with that of the macromolecule, the macromolecule can effectively be made to disappear in SANS. The *neutron match point* (Table 23.3) is the percentage D_2O in the solvent where scattering from macromolecule and solvent are equal.

Contrast variation can be used to study macromolecular complexes. The trick is to find a scattering density (H_2O/D_2O ratio) for the solvent that matches some part of a macromolecular complex and thus makes this part invisible to SANS so that the scattering from the complex corresponds only to the visible parts that do not match the solvent. For example, the layer of detergents that is used to solubilize a membrane protein can be made invisible to SANS by contrast variation. The SANS data then report only on the membrane protein itself, without the contribution of the detergent. Contrast matching is also applied in the distinction between proteins and nucleic acids, which have different neutron match points (Table 23.3). By matching the H_2O/D_2O ratio to either protein or nucleic acid, the relative location of the two components can be addressed. An example is the nucleosome: originally, it was unknown whether in nucleosomes (Figure 19.28) the DNA was wrapped around protein or whether protein would enclose the DNA. By matching the solvent selectively to either the protein or the DNA and comparing the R_g values from different SANS experiments, it became clear that the DNA was wrapped around the histone proteins (Figure 23.13).

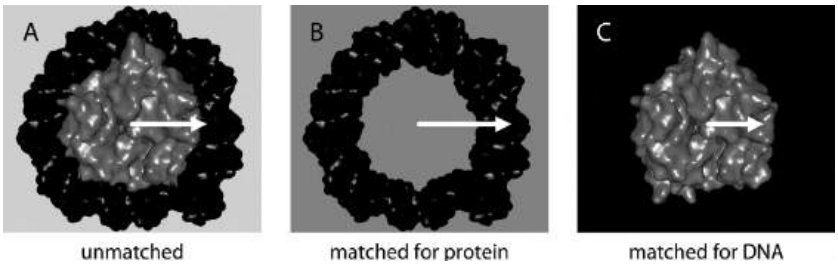


FIGURE 23.13: Principle of contrast variation in SANS. The nucleosome serves as an example to distinguish whether DNA is wrapped around proteins or *vice versa*. From left to right the scattering density of the solvent is increased, indicated by the light gray, dark gray, and black backgrounds. R_g, indicated by a white arrow, is the measured variable. The structures themselves are shown as a reference but are not obtained from SANS. A: The solvent is 100% H_2O with low scattering density, as is typically used in SAXS. SAXS data are therefore similar to SANS data collected in 0% D_2O, and both the DNA and the protein contribute to R_g. B: The solvent scattering density matches that of the protein (40–45% D_2O). SANS is dominated by the ring-shaped DNA, which leads to a larger R_g than the whole complex in (A). C: When the solvent scattering density matches that of DNA (65% D_2O) only the protein remains visible with a small R_g.

Contrast variation in SANS is also useful to gather distance information on macromolecular complexes. In the 1970s, before the crystal structure of the ribosome (see Figure 19.39) was known, the location of the constituent ribosomal proteins was established by SANS: deuterated ribosomes were isolated from bacteria grown in deuterated media. Individual ribosomal proteins with ^{1}H instead of D were incorporated singly or in pairs into the deuterated ribosomes (a rather complicated assembly process on its own), and SANS data were collected in D_2O. R_g and D_{max} for the ^{1}H substructures (the two proteins within the ribosome) were extracted from the distance distribution functions. Three SANS experiments on the singly and doubly modified ribosomes resulted in the R_g values for the individual proteins and the substructure, from which the distance between the two components was determined. Repeated SANS experiment with many different pairs of subunits yielded a triangulation map of the ribosomal proteins whose only uncertainty was its chirality, as in the *ab initio* SAXS calculations.

QUESTIONS

23.1 The Rayleigh ratio is given by

$$R_\theta = \frac{I_s}{I_0} \cdot \frac{r^2}{V \cdot \left(1 + \cos^2 \theta\right)}$$

The absorption is defined as the logarithm of the ratio of the incident and transmitted light intensities:

$$A = \log \frac{I_0}{I}$$

Thus, the apparent absorption due to scattering is

$$A_s = \log \frac{I_0}{I_0 - I_s}$$

Calculate A_s as a function of R_θ. Choose a suitable volume V and integrate over all angles θ to obtain the total scattering.

23.2 Raman scattering leads to two bands, the Stokes band and the *anti*-Stokes band. In fluorescence spectroscopy, the position of the Raman peak as a function of the excitation wavelength λ_{ex} is given by:

$$\lambda_{Raman} = \frac{\lambda_{ex}}{1 - 36 \cdot 10^{-5} \cdot \lambda_{ex}}$$

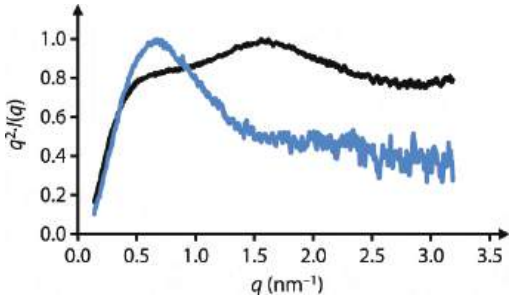
To which of the two Stokes bands does the Raman peak correspond, and why? You want to titrate a 1 μM solution of a protein with a high-affinity ligand and use Trp fluorescence as the spectroscopic probe. The protein solution absorbs maximally at 287 nm and emits fluorescence maximally at 320 nm. Would you excite at the absorption maximum? Which excitation wavelength would place the Raman peak 10 nm red-shifted from the fluorescence maximum? Are such considerations important when using a 10-fold higher concentration of protein?

23.3 At a given scattering angle θ, what is the relative scattering intensity of argon atoms Ar, chloride ions Cl^-, and calcium ions Ca^{2+}? What about H_2O, F^-, and Mg^{2+}?

23.4 Calculate the average electron density of water.

23.5 Derive a formula that relates the molecular mass M_M of a spherical protein (in Da) to the radius of gyration R_g (in nm). The partial specific volume of proteins is 0.73 $cm^3\ g^{-1}$.

23.6 The figure shows the Kratky plots of an RNA molecule, the regulatory domain of the Lys riboswitch, both in the absence and in the presence of 5 mM $MgCl_2$ (all other buffer conditions are identical). What is the explanation for the different curve shapes?



23.7 What are the frequency, temperature, velocity, and wavelength of neutrons with energy of 25 meV? One eV equals $1.602 \cdot 10^{-19}$ J, the Planck constant is $h = 6.62607 \cdot 10^{-34}$ J s, and the neutron mass is $m_n = 1.67493 \cdot 10^{-27}$ kg.

REFERENCES

Aydin, O., Altas, M., Kahraman, M., Bayrak, O. F. and Culha, M. (2009) Differentiation of healthy brain tissue and tumors using surface-enhanced Raman scattering. *Appl Spectrosc* **63**(10): 1095–1100.
· use of surface-enhanced Raman scattering to distinguish between healthy and tumor cells

Cotton, T. M., Kim, J. H. and Chumanov, G. D. (1991) Applications of surface-enhanced Raman spectroscopy to biological systems. *J Raman Spectr* **22**: 729–742.
· review article that summarizes the application of Raman spectroscopy to small molecules, DNA, soluble and membrane proteins, and membranes

Cotton, T. M., Schultz, S. G. and Van Duyne, R. P. (1980) Surface-enhanced resonance Raman scattering from cytochrome c and myoglobin adsorbed on a silver electrode. *J Am Chem Soc* **102**(27): 7960–7962.
· surface-enhanced resonance Raman scattering

Franke, D., Jeffries, C. M. and Svergun, D. I. (2015) Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra. *Nat Methods* **12**(5): 419–422.
· use of correlation maps to judge the quality of models in *ab initio* modeling of SAXS curves

Hu, J. and Zhang, C. Y. (2012) Single base extension reaction-based surface enhanced Raman spectroscopy for DNA methylation assay. *Biosens Bioelectron* **31**(1): 451–457.
· surface-enhanced Raman spectroscopy to determine the degree of DNA methylation

Hudson, S. D. and Chumanov, G. (2009) Bioanalytical applications of SERS (surface-enhanced Raman spectroscopy). *Anal Bioanal Chem* **394**(3): 679–686.
· review on bioanalytical applications of surface-enhanced Raman spectroscopy

Jeffries, C. M. and Svergun, D. I. (2015) High-throughput studies of protein shapes and interactions by synchrotron small-angle X-ray scattering. *Methods Mol Biol* **1261**: 277–301.
· synchrotron small-angle X-ray scattering

Kozin, M. B. and Svergun, D. I. (2001) Automated matching of high- and low-resolution structural models. *J Appl Cryst* **34**(1): 33–41.
· introduces the normalized spatial discrepancy as a means to compare structures not represented by atomic coordinates but by an array of points

Lewis, A., Fager, R. S. and Abrahamson, E. W. (1973) Tunable laser resonance Raman spectroscopy of the visual process. *J Raman Spectr* **1**: 465–470.
· resonance Raman spectroscopy of rhodopsin

Miura, T., Benevides, J. M. and Thomas, G. J., Jr. (1995) A phase diagram for sodium and potassium ion control of polymorphism in telomeric DNA. *J Mol Biol* **248**(2): 233–238.

· use of different Raman spectral signatures for parallel and antiparallel G-quadruplex structures to analyze the populations in telomeric DNA in equilibrium

Palacky, J., Vorlickova, M., Kejnovska, I. and Mojzes, P. (2013) Polymorphism of human telomeric quadruplex structure controlled by DNA concentration: A Raman study. *Nucleic Acids Res* **41**(2): 1005–1016.
· characterization of the concentration dependence of G-quadruplex polymorphism by Raman spectroscopyPerkins, S. J. (1988) *Modern Physical Methods in Biochemistry, Part B*. Elsevier.

· scattering densities and electron densities for biological substances

Putnam, C. D., Hammel, M., Hura, G. L. and Tainer, J. A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution. *Q Rev Biophys* **40**(3): 191–285.
· a very informative review on the combination of SAXS and X-ray crystallography that highlights the close physical relationship between the two methods

Spiro, T. G. (1974) Resonance Raman Spectroscopy: A new structure probe for biological chromophores. *Acc Chem Res* **7**: 339–344.
· review article highlighting the principle of resonance Raman spectroscopy and its applications to biomolecules

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. and Zaccai, G. (1998) Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proc Natl Acad Sci U S A* **95**(5): 2267–2272.
· solvation of proteins can be detected by both SAXS and SANS, with the first hydration layer being about 10% more dense than bulk water

Talari, A. C. S., Evans, C. A., Holen, I., Coleman, R. E. and Rehman, I. U. (2015) Raman spectroscopic analysis differentiates between breast cancer cell lines. *J Raman Spectrosc* **46**: 421–427.
· original article demonstrating differences in lipid, nucleic acid, and protein composition of normal and malignant cell lines leading to different Raman fingerprints

Tuma, R. (2005) Raman spectroscopy of protein: From peptides to large assemblies. *J Raman Spectrosc* **36**: 307–319.
· review article with a focus on assignment of Raman resonances and applications to proteins

Tuukkanen, A. T. and Svergun, D. I. (2014) Weak protein-ligand interactions studied by small-angle X-ray scattering. *FEBS J* **281**(8): 1974–1987.

SAXS can be used to analyze the change in protein oligomerization state upon ligand binding, and low-affinity ligands amounting to 10–15% of the protein mass can be tracked directly by SAXS
Weidinger, I. M. (2015) Analysis of structure-function relationships in cytochrome c oxidase and its biomimetic analogs *via* resonance Raman and surface enhanced resonance Raman spectroscopies. *Biochim Biophys Acta* **1847**(1): 119–125.
· review article on the application of resonance Raman spectroscopy to cytochrome c oxidase and other heme-copper oxidases

Young, G., Hundt, N., Cole, D., Fineberg, A., Andrecka, J., …, Kukura, P. (2018) Quantitative mass imaging of single biological macromolecules. *Science* **360**(6387): 423–427.
· benchmarking study for mass photometry

24 Crystallography

The scattering techniques described in Chapter 23 give information on molecules in solution but are somewhat limited in their resolution, particularly because of the small and isotropic scattering signal of individual molecules. Structural information with atomic resolution is provided by crystallographic techniques, which measure the scattering of electromagnetic radiation by ordered arrays of molecules known as crystals. Since the maximum resolution attainable is determined by the wavelength of the radiation ($> \lambda/2$; Section 25.1), any radiation that interacts with crystalline matter and has a wavelength smaller than the covalent bond distances in molecules (100–250 pm) can in principle provide atomic resolution information. Useful radiation includes particles such as electrons and neutrons, but most frequently X-rays are used to probe crystals. Each of these probes have spawned their own field in crystallography. While X-rays and electrons interact with the electron clouds of atoms, neutrons interact with the nuclei of the sample. However, the physical basis behind scattering of electromagnetic radiation and particles on ordered objects, known as *diffraction*, is the same (Section 24.4.2).

24.1 PHASE PROBLEM AND REQUIREMENT FOR CRYSTALS

Construction of images from diffraction data requires both intensity and phase information, for which in optical microscopy lenses are used (Chapter 25). Unlike electrons that can be focused using magnetic lenses, no efficient lenses exist for radiation of short wavelengths, including neutrons and X-rays. All that can be done in such experiments is to collect the scattered light intensity as a function of scattering angle, but, in doing so, the relative phase information between the scattered beams is lost. While acceptable for low-resolution information on molecules, as in SAS (Section 23.2), high-resolution information on atomic positions absolutely requires phases. This issue is known as the *phase problem*. In crystallography, the image of the molecule is reconstructed from the intensity measurements using *phasing methods* (Section 24.4.4) and mathematical transformations that include the Fourier transformation (Section 31.15).

Another severe problem for high-resolution scattering studies is signal strength. The scattered light intensity given off by a single macromolecule is so tiny that currently not even the most intense X-ray sources, the free electron lasers (Section 24.6), are strong enough to retrieve structural information from single molecules. This problem cannot simply be solved by increasing the concentration

of particles, since molecules in solution are not ordered and hence scatter in all directions, which leads to weak signals. By arranging molecules into crystals, we produce an amplifier for the scattering signal. Because all molecules have defined relative orientations, their scattering follows defined phase relationships. The scattering from a crystal (Section 24.4.2) is very different from that of a molecule in solution: the scattering intensity from crystals is limited to certain diffraction angles, so the total diffraction intensity concentrates in certain areas and can accurately be measured to high resolution. An important step that lies before the determination of a *crystal structure* is therefore the generation of well-ordered crystals.

24.2 CRYSTALLIZATION OF MACROMOLECULES

A crystal is a regular two- or three-dimensional arrangement of objects (Figure 24.1). Macromolecular crystals are held together by weak non-covalent interactions between surface atoms of neighboring molecules, particularly hydrogen bonds, electrostatic interactions, and van der Waals interactions (see Section 17.2.2). The percentage of surface areas that provide these *crystal contacts* becomes smaller and smaller for larger molecules: if we consider a sphere of radius r, the volume scales with r^3, but the surface scales only with r^2. Consequently, larger macromolecules can form fewer crystal contacts than smaller macromolecules and tend to be more difficult to crystallize. The forces acting between the molecules are actually smaller than the forces exerted during conformational changes. For instance, ligand binding to a crystalline protein frequently causes the crystals to crack if the associated conformational changes are large. A macromolecule is anything but "fixed" in a crystal.

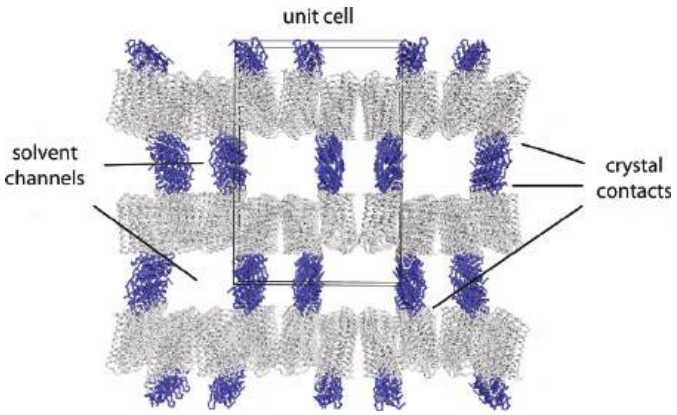


FIGURE 24.1: Macromolecular crystal. A protein-protein complex between a nanobody (blue) and a GPCR (gray) forms a regular arrangement. The nanobody serves as a *crystallization helper* by acting as spacer between the GPCR molecules. Large solvent channels are visible between the proteins.

As a result of the few crystal contacts, large solvent-filled voids are present between the macromolecules that form a network of *solvent channels* running through the crystal (Figure 24.1). The fractional volume not occupied by macromolecules is the *solvent content* of a crystal. On average, the solvent content is 50% of the total crystal volume, but crystals with solvent contents from 25% to 90% have been characterized. Macromolecular crystals therefore resemble an ordered gel with large pores, quite unlike the hard small molecule crystals of sugar and table salt, which are densely packed and contain no solvent. Together with the few and weak interactions stabilizing macromolecular crystals, their high amount of water renders them quite fragile: crystals break easily during handling, change of temperature, or slight dehydration. They also tend to age, suffering from air oxidation and slow chemical reactions with the crystallization components. Absorption of X-rays or electrons by the solvent in macromolecular crystals produces radicals that induce *radiation damage*, a severe problem in crystallography that can be overcome partly by cryo-cooling (see Box 24.2).

The crystallization process can be conceptually divided into a kinetic and a thermodynamic part. A crystal will form only if the overall ΔG of the assembly is negative, which is equivalent to stating that molecules in a solution of certain composition (at supersaturation) have a higher free energy (Section 1.7.1) than the crystal. Whether the crystal forms on a reasonable time scale depends on the activation energy that is associated with forming a critical nucleus (Figure 24.2), a tiny, ordered assembly capable of growing into a macroscopic crystal. Nucleus formation is reversible, and the smaller the nucleus, the more likely it is to dissolve again. A competing pathway to crystallization is aggregation, which is usually irreversible. Both aggregation and nucleus formation are favored by high concentrations of macromolecules and thus can occur at the same time under the same conditions. Successful crystallization of macromolecules is therefore a delicate balance between these competing pathways.

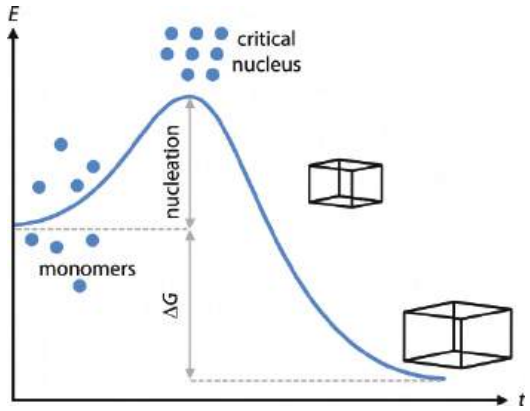


FIGURE 24.2: Energy diagram of crystallization. Under conditions of supersaturation, molecules in solution represent a higher energy state than the crystal. Both states are separated by an activation energy barrier (nucleation). The ordered nucleus grows into a crystal until the solid phase is in equilibrium with the solution phase.

Unfortunately, it is impossible to predict the crystallization conditions from general properties such as sequence, isoelectric point, or molecular mass. Crystallization conditions may radically change even for proteins and nucleic acids that differ only by a single residue or that have been purified by different methods. Consequently, crystallization is a process of trial and error, and a lot of creativity goes into the design of crystallizing samples and the development of crystallization methods. Typically, several hundred to a few thousand crystallization trials are required until ordered crystals are found or it is judged that the macromolecule to be crystallized needs to be modified. The whole process can last from days to years and consume anything between just 1 mg or several grams of sample.

The techniques used for crystallization include *dialysis*, *free interface diffusion*, *micro-batch*, *vapor diffusion*, and *seeding* (Figure 24.3). All crystallization methods use the same strategy. The macromolecular solution at 0.5–50 mg mL^{-1} concentration is mixed with a *precipitant* to bring the solution into a supersaturated state from which crystals might grow. Useful precipitants are salts (0.5–4M), organic solvents (ethanol, propanol, MPD or 2-methyl-2,4-pentanediol at 5–25%), and polyethylene glycol (PEG; 5–50%) of average molecular mass between 200 Da and 20 kDa. Other parameters that can be adjusted are pH (usually between 3 and 10) and temperature (4–40°C).

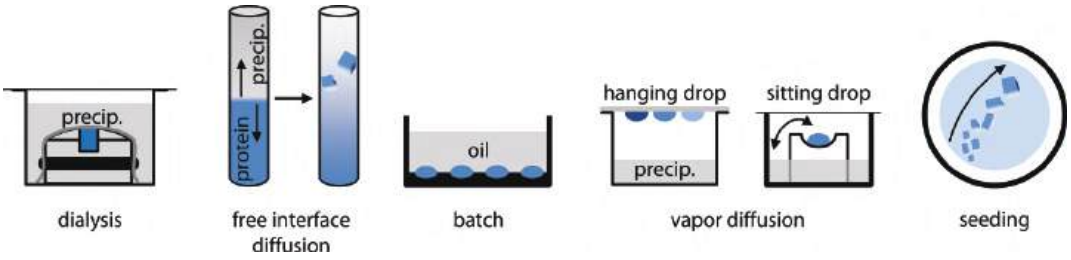


FIGURE 24.3: Crystallization methods. The methods are shown from left to right according to a decreasing amount of sample that is usually needed. Protein solution is shown in blue color, and the precipitant (*reservoir*) is shown in gray. In dialysis, the protein is separated from the precipitant by a membrane, while free interface diffusion does not need a membrane. In batch crystallization the protein is mixed with precipitant and placed under oil. Vapor diffusion comes in two variations, hanging and sitting drop. Several hanging drops, *e.g.* different protein concentrations or ligand complexes, can be placed over the same reservoir. Seeding can be applied to any of the techniques. The direction of the arrow marks the trajectory of a hair dipped into a seed solution and dragged through a sitting drop. The seeds are diluted along the trajectory, leading to fewer crystals at its end. Batch and sitting drop vapor diffusion have been miniaturized into 96-well formats.

For dialysis, a protein solution is placed in a *dialysis button* of 5–50 µL volume, closed off by a semipermeable membrane, and equilibrated against a reservoir of precipitant. The rates of water and ion fluxes through the membrane are determined by the pore size of the membrane and the concentration gradients across it. Free interface diffusion is carried out in narrow capillaries where mass transport occurs predominantly by diffusion but not convection. The macromolecule and precipitant solutions are brought into direct contact (free interface). Diffusion of precipitant into the macromolecule solution and *vice versa* creates concentration gradients of opposite directions. At some point along the capillary, the system may reach supersaturation and form crystals. In microbatch crystallization, supersaturation is achieved quasi instantaneously by mixing ca. 0.5–2 µL of a macromolecular solution with a precipitant in a 1:3 to 3:1 ratio, depending on the solubility of the macromolecule. The droplet is then sealed with a layer of oil. Probably the most popular crystallization method is vapor diffusion. Droplets of 0.02–1 µL volume are prepared by mixing macromolecule and precipitant solutions in various ratios (1:2, 1:1, 2:1, *etc.*), and equilibrating them against a 0.05–1 mL *reservoir* of the precipitant solution. After mixing with reservoir, the droplet has different concentrations and hence different chemical potential (Section 1.7) of all components and therefore will exchange volatile substances (water, alcohols) with the reservoir. This exchange is mostly from the droplet to the reservoir (the drop shrinks) but can also be reversed, for example if the protein buffer is of higher osmolarity than the precipitant solution. In the hanging drop method, droplets on a cover slip are suspended over the reservoir while in the sitting drop method, the droplet rests on a post above the reservoir solution. Usually carried out in 96- or 384-well plates by a robot, the sitting drop method is ideal for large-scale screening of hundreds of different crystallization conditions. Many sitting-drop crystallization plates have several depressions next to the reservoir that allow variations of the sample such as different macromolecule concentrations or ± ligand when testing against the same precipitant solution.

The different crystallization methods achieve supersaturation in different ways, and they follow different paths in the phase diagram toward the equilibrium (Figure 24.4). A crystallization trial is basically a search for the region in the phase diagram where the crystalline phase is stable.

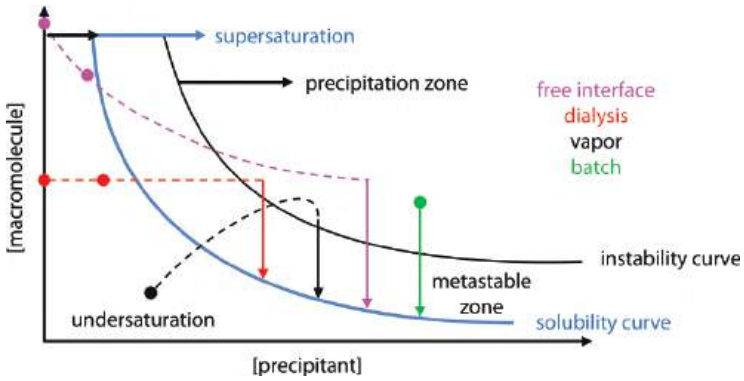


FIGURE 24.4: Schematic crystallization phase diagram. The solubility curve (blue line) separates the undersaturation and supersaturation zones. The supersaturation zone is subdivided into a metastable and a precipitation zone. Crystals form in the precipitation zone and continue to grow in the metastable zone. Four crystallization methods and the routes they take to arrive *via* the nucleation zone at the solubility curve are highlighted in color. Dots represent possible starting concentrations of macromolecule and precipitant. Free interface diffusion and dialysis may start with premixed precipitant and protein. In free interface diffusion, every part of the capillary follows a different path. Figure modified after Chayen (1998). Acta Cryst. D 54: 8–15 and reproduced with permission of the IUCR.

The *solubility curve* in the phase diagram represents the solubility of a macromolecule as a function of precipitant concentration (Figure 24.4). At lower precipitant concentrations is the stable zone of undersaturation, where the maximum protein concentration is not reached at the given precipitant concentration. Any crystals that might be present in this zone would dissolve. The supersaturation zone to the right of the solubility curve is conceptually divided into a *metastable zone* and a *precipitation zone* by the *instability curve*. The likelihood of nucleation is thought to increase from the solubility curve toward the instability curve. Nucleation can occur in both the metastable and the precipitation zone. The metastable zone will at some point separate into a protein-rich phase (aggregate, droplets, or crystals), while the system to the right of the instability curve spontaneously decomposes into precipitate (again aggregate, droplets, or crystals) and saturated macromolecule solution. Once nuclei have formed, the protein concentration drops, and the system linearly traverses the metastable zone toward the solubility curve.

The formation of a nucleus in the supersaturation zone can be the rate-limiting step for crystallization (Figure 24.2). This step can be skipped by introducing seeds that function as external nuclei. Seeds are usually made by smashing unusable crystals, either of the same or of another protein. It is not uncommon for seeds to work in very different conditions compared to where they originated from (*cross-seeding*). Alternatively, small mineral crystallites may be used, often serendipitously introduced in the form of dust particles. Seeds may be introduced by pipetting or with a smooth hair (historically a cat whisker or horsehair) dipped into a seed suspension and dragged through the crystallization drop. Using a hair dilutes the seeds attached to it and often leads to a shower of crystals along the trajectory with some larger crystals at the end (Figure 24.3).

Due to the many parameters that can be adjusted, searching the crystallization phase diagram for the crystallization zone is a multi-dimensional problem. In addition, the likelihood of crystallization also depends on properties of the sample: its purity, entropy, and surface hydrophobicity. In general, a pure sample of rigid molecules and polar surfaces is less likely to aggregate and will more likely crystallize than an impure sample of flexible and hydrophobic molecules. Impurities can bind to growing crystals and poison their growth by interrupting the regular pattern of the molecules. High entropy of the molecules in the sample translates into conformational heterogeneity, which reduces the likelihood of incorporation into the crystal. Very hydrophobic molecules also tend to aggregate faster than they form crystals. Two general strategies to increase the likelihood of crystallization are therefore reduction of the entropy and reduction of the amount of solvent-exposed hydrophobic surfaces.

Large multi-domain proteins with flexible linkers have high entropy and thus do not crystallize easily. By contrast, their individual domains, often identified by limited proteolysis, tend to crystallize more easily (*divide et impera* approach). Limited proteolysis of termini and loops is a general method to reduce the surface entropy of a protein by removing flexible parts. Amino acids with long side chains (Lys, Arg, and Glu), contributing to high surface entropy, can also be mutated to smaller amino acids. This *surface entropy reduction* (SER) requires some prior knowledge on which side chains are likely to be on the surface, a task that is greatly facilitated by looking at AlphaFold models (Section 20.3.2). By contrast, proteins from thermophiles naturally have reduced surface entropy compared to their mesophilic and psychrophilic counterparts: on average, thermophilic proteins have shorter loops and less flexible side chains at their surface, and the reduced entropy contributes to their increased thermostability. Structures of orthologs are similar at the core but different at the surface where the non-conserved residues cluster, so choosing a thermophilic homolog is a natural way to exploit SER for crystallization.

Finally, if variation of the protein itself is unsuccessful, *crystallization helpers* may be used (Figure 24.5). Crystallization helpers are very stable and readily crystallizing proteins that bind with high affinity to hydrophobic areas or flexible loop regions, thus reducing the hydrophobicity and/or entropy of the sample and increasing the likelihood of crystallization. This is especially useful for the crystallization of membrane proteins (Box 24.1) and flexible RNA (Figure 24.5). Crystallization helpers include antibody Fab, nanobodies, DARPins (designed ankyrin repeat proteins), and the U1A protein.

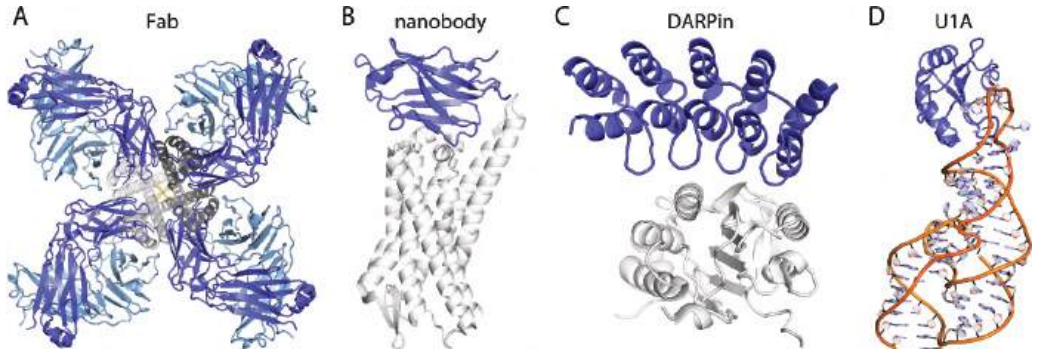


FIGURE 24.5: Helper proteins for crystallization. Helper proteins (blue) are often responsible for the majority of intermolecular contacts in a crystal. A: Antibodies have been raised against the voltage-gated K^+ channel KcsA and the antigen-binding fragments (Fab; light and dark blue) were used to generate a crystallizing complex (PDB-ID 4uuj). As the K^+ channel is a homotetramer, there are four Fabs bound in this complex. B: The μ-opioid receptor is a GPCR that in this example has a llama nanobody bound (blue; PDB-ID 5c1m). C: DARPins have an array of loops that can be engineered to bind to any epitope, as in this example of caspase–7 (PDB-ID 4jb8). D: The U1A protein binds to a conserved RNA tetraloop that was engineered into the hepatitis δ virus ribozyme (PDB-ID 1drz), a very successful method to crystallize RNA molecules in general.

Antibodies can be raised against both proteins and nucleic acids, and limited proteolysis is used to generate the Fab. A smaller version of Fab is nanobodies, which contain only a single immunoglobulin domain and are derived from camelid or shark antibodies. Like antibodies, DARPins have loop regions of variable sequence that can be generated to bind to virtually any protein or nucleic acid.

BOX 24.1: CRYSTALLIZATION OF MEMBRANE PROTEINS.

Membrane proteins are more difficult to crystallize than soluble proteins. Once removed from their lipid environment, membrane proteins are unstable and tend to aggregate (Section 18.4). Much effort goes into the stabilization of amphiphile-solubilized membrane proteins. Alkyl maltosides and glycosides were the amphiphiles used in about half of all membrane protein structure determinations. Random mutagenesis of the membrane protein followed by a test for improved thermostability (see Box 18.8) is also employed, but it may take a long time until a construct is obtained that is stable enough to allow crystallization. Lastly, binding of the crystallization helper proteins that mask hydrophobic surface areas and/or reduce aggregation is becoming increasingly popular (Figure 24.5). If these attempts fail, membrane proteins may be crystallized from a *lipidic cubic phase* (LCP, *in meso* crystallization) that mimics a membrane-like environment (Figure 24.6). LCP are bicontinuous crystalline (*i.e.* repeating) phases not unlike a sponge that form spontaneously upon mixing of suitable lipids with water. Examples of such lipids are the mono-glycerides monoolein and monopalmitolein and the di-terpene alcohol phytantriol. The *mesophases* thus obtained are mixed with the detergent-solubilized membrane protein and serve as crystallization matrices within which the proteins can diffuse laterally and assemble into a crystal. The crystals obtained from membrane proteins, particularly those of higher eukaryotes, are usually tiny, extremely fragile, and sensitive to any change in buffer composition.

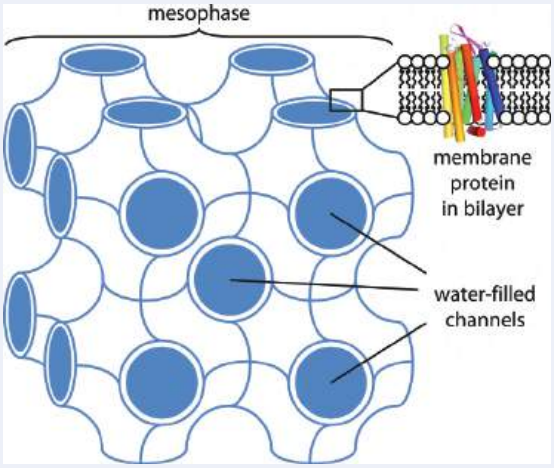


FIGURE 24.6: Lipidic cubic phase. Lipids such as monoolein and monopalmitolein can form membranes that further organize into a mesophase, *i.e.* a three-dimensional structure with water-filled tubes and channels. Membrane proteins solubilized by detergents may concentrate in the membranes of the mesophase and crystallize.

Crystallization of nucleic acids has the general problem of electrostatic repulsion between the phosphate backbones of molecules, so counter ions are essential for charge neutralization. Most of the contacts in nucleic acid crystals are made *via* the phosphate backbone and metal ions, but specific base-mediated contacts, such as kissing loops (Section 19.6.1), can also be essential for crystallization. Kissing loops may be engineered into RNA to promote crystal formation. Alternatively, a 10 base pair stem-loop structure that is recognized by the 11 kDa RNA-binding domain of the spliceosomal U1A protein can be included in RNA. The biophysical properties of protein surfaces are more variable than those of nucleic acids, so the inclusion of U1A has become a popular technique in RNA crystallization (Figure 24.5). Because the isoelectric point of U1A is > 10, electrostatic interactions in addition to protein-RNA interactions can promote crystallization of RNA/U1A complexes.

BOX 24.2: X-RAY RADIATION DAMAGE AND CRYO-PROTECTION.

In crystallography, the term *radiation damage* describes the detrimental effects of absorbed X-rays on a crystal. The first step in radiation damage is photoelectric absorption whereby an X-ray photon transfers its energy to an electron, ejecting it from its atomic orbital (see Figure 24.12). At room temperature this *photoelectron* can ionize hundreds of other atoms, producing a few hundred more electrons in the process. The atom that is left behind without the electron may return to its ground state by emission of an X-ray photon (see Figure 24.12) or it may emit yet another electron, the *Auger electron*. All these electrons traveling through the crystal will produce radical species, mostly from water ($OH^{\bullet}$ and $H^{\bullet}$). Radicals and electrons induce bond breakage by chemical reactions either directly with the surrounding macromolecules or, after diffusion through the crystal, with distant macromolecules. The result of radiation damage is a lower resolution crystal structure (*global damage*) with less information content than it could potentially have (*specific damage*). While global damage is a general reduction in diffraction intensity, specific damage includes reduction of disulfide bonds to Cys and of metal ions in metalloproteins (Fe^{3+} to Fe^{2+}, for example), decarboxylation of acidic side chains, de-hydroxylation of Tyr, and destruction of aromatic rings, particularly in nucleic acid crystals. At room temperature, radiation damage sets in within about 100 ms after exposure of a crystal with synchrotron light. The maximum dose (energy per unit mass) tolerated by a crystal is only 150 kGy, or 150 kJ of radiation energy per kg of matter.

The extent of radiation damage is much reduced if data collection is done at low temperature where the photoelectrons and radicals generated by X-rays diffuse more slowly. This leads to a dramatic increase of crystal lifetime in the X-ray beam. The first structures using cooling were determined at 4°C, well above the freezing point of the crystallization mother liquor. Most crystal structures are nowadays determined at cryogenic temperatures by placing the crystal in a stream of gaseous nitrogen at its boiling point (about 100 K). At this temperature, the crystallization mother liquor forms a solid, effectively abolishing diffusion within the crystal. The problem of this *cryo-cooling* is that water in the mother liquor may eventually freeze into ice crystals that inevitably have a larger volume than bulk water, destroying the macromolecular crystal. *Cryo-protection* of macromolecular crystals seeks to avoid water ice formation and to reduce the expansion of the mother liquor upon cryo-cooling as much as possible. The standard way to inhibit ice formation is to disrupt the hydrogen-bonded water clusters using cryo-protectants that compete with water for hydrogen bonds. These include polyethylene glycol (PEG), low molecular mass alcohols (ethylene glycol, glycerol, 2,3-butanediol, 2-methyl-2,4-pentanediol, 1,6-hexanediol, glucose, xylitol, saccharose) or cryosalts ($\geq$ 1 M Li^{+} salts or the organic salts malonate, formate, citrate, tartrate, and acetate). Cryo-protection and data collection at 100 K increases the maximum dose tolerated by a crystal to ca. 20 MGy. Once this *Henderson limit* is reached, the diffraction pattern of the crystal has declined to half the original intensity.

24.3 SYMMETRY AND SPACE GROUPS

A basic understanding of crystal symmetry is useful to better understand the process of diffraction. The regular arrangement of objects in a crystal brings two main characteristics: periodicity and symmetry. The periodicity of crystals is described by their *lattice*, and their symmetry is determined by the *space group*. In practice, the lattice and space group are determined from the measured diffraction data.

If we cut a crystal in two pieces, we obtain two smaller crystals of the same periodicity and symmetry. We can repeat this Gedankenexperiment until we arrive at the smallest conceivable "crystal",

the *unit cell*. The unit cell is a parallelepiped, defined by three *crystal axes a*, *b*, and *c* that enclose three angles of magnitude α, β, and γ. The crystal axes form a right-handed coordinate system represented by three unit vectors ***a***, ***b***, and ***c*** (Figure 24.7). The magnitudes of the unit vectors are the *unit cell constants*, typically between 3 nm and 300 nm for macromolecular crystals.

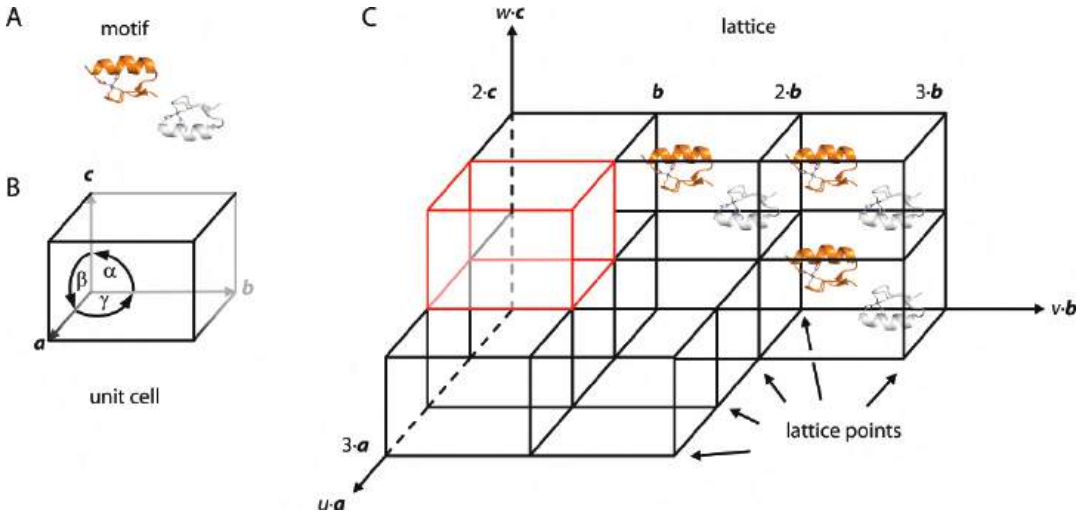


FIGURE 24.7: Unit cell and crystal lattice. A: The motif or asymmetric unit that builds up the unit cell by application of symmetry operators. In this case, there are two asymmetric units per unit cell. B: The unit cell is defined by three axes lengths and three angles. The *a*- and *b*-axes enclose the γ-angle, the *b*- and *c*-axes enclose α, and the *c*- and *a*-axes enclose β. Mathematically, the unit cell is described by three basis vectors ***a***, ***b***, and ***c***. C: The macroscopic crystal is a result of three-dimensional translation of the unit cell. Three of the unit cells are filled with their asymmetric units. The lattice points (arrows) are located at the intersections of the lines. Unit cells that have lattice points only at their corners as in this example build up a primitive lattice.

Translation of the unit cell along its edges in all three dimensions builds up the macroscopic crystal. The crystal lattice is a mathematical concept that evenly divides space by sets of parallel and equidistant planes. The lattice is infinite in all three dimensions and covers all available space without gaps. The requirement to fill space completely means that only lattices with 1-fold, 2-fold 3-fold, 4-fold, and 6-fold rotational symmetries are possible. For example, 6-fold rotational symmetry means that after rotation of the lattice by 360°/6 = 60° the same three-dimensional pattern of lattice points is achieved. The different rotational symmetries give rise to the seven *crystal systems* triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic (Figure 24.8). These crystal systems not only differ among themselves by their internal symmetry but also by the requirements for their unit cell constants, *i.e.* the relative length of the axes and the angles enclosed by the unit cell vectors (Table 24.1). For example, the cell dimensions of a triclinic unit cell have no restrictions, and it has 1-fold symmetry (*i.e.* none): only a full rotation by 360° will superimpose the lattice onto itself. By contrast, cubic cells must have all axes of the same length, all angles must be 90°, and there are four 3-fold symmetry axes along the body diagonals of the cube (Table 24.1). It is important to note that the variable parameters, indicated by a "≠" in Table 24.1, may be different but do not necessarily have to be different from the fixed parameters. It is entirely possible that by diffraction measurement a triclinic lattice is found with cell dimensions that resemble that of a cube. The cell dimensions determined from diffraction data (Section 24.4.3) are a hypothesis that is only validated once the crystal structure has been determined.

At the intersections of the lattice planes lie the *lattice points*. All lattice points are equivalent because they have the same environment. The different lattice symmetries in three-dimensional crystals are abbreviated P, C, I, and R. Primitive (P) lattices have lattice points only at the corners of the unit cells (Figure 24.7). The lattice points are shared by eight adjoining unit cells, so there is a total of one lattice point per unit cell. Lattices other than primitive have additional lattice points.

Single face-centered, or C-centered, lattices have unit cells with an additional lattice point at the center of one pair of opposing faces. In face-centered, or F-centered lattices, all three pairs of opposing faces have a lattice point at the center of each face. I-centered (body-centered) lattices have an additional lattice point at the intersection of their body diagonals. Lastly, the rhombohedral R-lattice is an alternative description of a hexagonal lattice. Not all combinations of lattice types are possible with all crystal systems. For example, the triclinic crystal system is only compatible with a primitive lattice. In total, there are fourteen *Bravais lattice symmetries* (Figure 24.8).

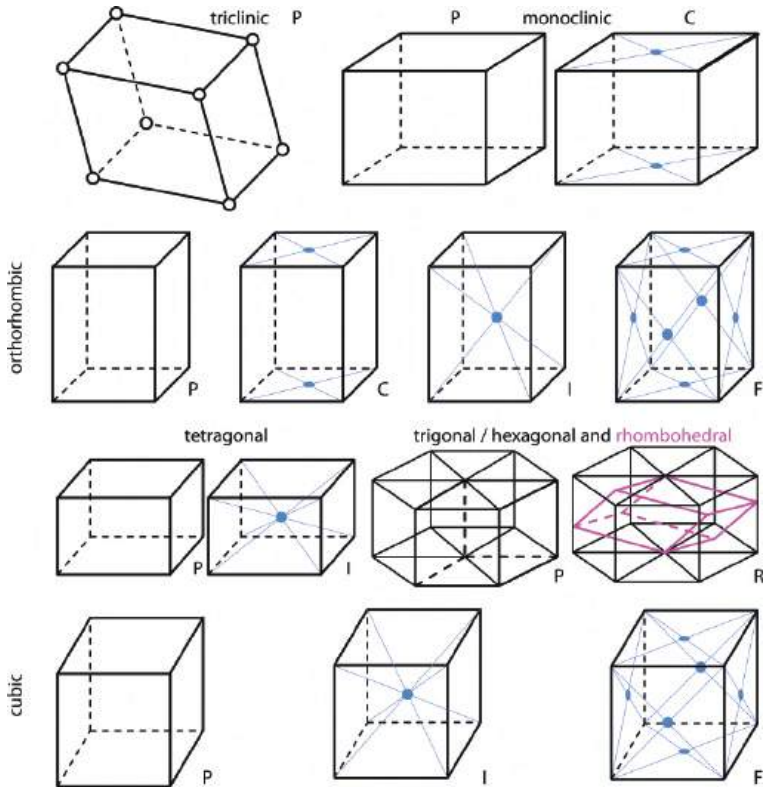


FIGURE 24.8: The fourteen Bravais lattices and seven crystal systems. Each parallelepiped is the building block of a Bravais lattice. The crystal systems are trigonal, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic. All of these can have a primitive (P) lattice. C, I, and F are the C-centered, face-centered, and body-centered lattice types, which are not possible for all crystal systems. The additional lattice points in C, I, and F lattices are marked by blue dots. C- and I-lattices have one additional lattice point, and F-lattices have three additional lattice points. A rhombohedral cell (R-type lattice) is a cube stretched or compressed along one of its body diagonals. Rhombohedral cells can be described by a hexagonal cell three times as large, so the rhombohedral system is part of the hexagonal system. You can construct your own crystal systems by folding the two-dimensional cutouts provided at periodni.com/download/models_of_crystal_systems.pdf.

If we further dissect the unit cell, we arrive at the motif or *asymmetric unit*. The asymmetric unit is the smallest entity from which the unit cell can be built by application of *symmetry operators*, which are mathematical operations that describe translations, rotations, mirror and inversion symmetry, and combinations thereof. An example of a symmetry operator is $(-x,-y,z+½)$, a short-hand notation for a two-fold rotation of a point (x,y,z) about the z-axis, followed by a translation of ½ its length along the z-axis:

$$\left(-x,-y,z+\frac{1}{2}\right)=\boldsymbol{R}\cdot\boldsymbol{x}+\boldsymbol{T}=\begin{pmatrix}-1&0&0\\0&-1&0\\0&0&1\end{pmatrix}\cdot\begin{pmatrix}x\\y\\z\end{pmatrix}+\begin{pmatrix}0\\0\\0.5\end{pmatrix} \qquad \text{eq. 24.1}$$

R is the *rotation matrix* operating on the vector $\boldsymbol{x} = (x,y,z)$, and ***T*** is the *translation vector*. The combination of this rotation and translation results in a screw axis, in this case a 2_1-axis (Figure 24.9). In general, an a_b screw axis rotates a point by 360°/a about an axis, and then applies a translation of b/a along that axis. A 2_1-axis is therefore a 180° rotation, followed by a shift of ½ along the axis.

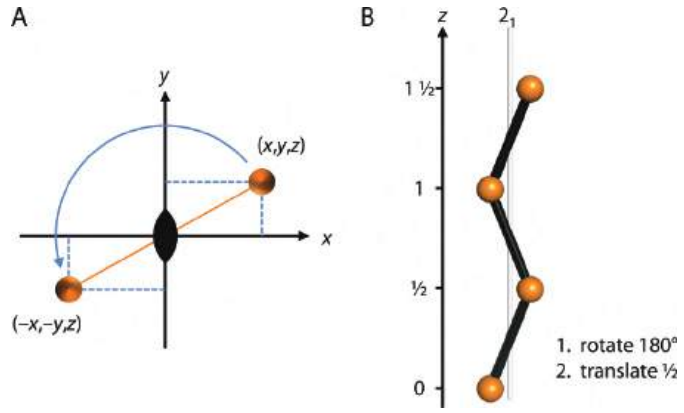


FIGURE 24.9: Rotational symmetry. A: View along the *z*-axis. The two-fold rotation (indicated by the lens-shape) about the *z*-axis converts coordinates (x,y,z) into $(-x,-y,z)$, which is equivalent to the rotation matrix in eq. 24.1. B: Side-view from (A). The translation by ½ of the *z*-axis length along the *z*-axis is encoded in the translation vector $\boldsymbol{T} = (0,0,½)$ in eq. 24.1.

There are only 230 different ways to arrange objects in three dimensions so that the same pattern is repeated infinitely in all directions. These *space groups* combine the information of a Bravais lattice (P, C, I, F, R) with the internal symmetry of the unit cell. Each space group is a complete description of the symmetry of the entire crystal by its characteristic set of symmetry operators. A limitation comes into play with the chirality of biological (macro)molecules. In the vast majority of cases, the biologically active isomer is a single enantiomer. Omitting all space groups containing symmetry operators that change the handedness of a molecule (mirror and inversion symmetry) limits the number of space groups accessible for chiral molecules to the 65 *chiral space groups* listed in Table 24.1.

TABLE 24.1: Crystal systems, Bravais lattices, and chiral space groups.

system	Bravais type	minimum internal symmetry	unit cell properties	chiral space groups
triclinic	P	none ("1-fold axis")	$a \neq b \neq c$, $\alpha \neq \beta \neq \gamma$	P1
monoclinic	P, C	2-fold axis ‖ ***b***	$a \neq b \neq c$, 90, $\beta \neq 90$, 90	$P2_{(1)}$, C2
orthorhombic	P, C, I, F	three 2-fold axes, mutually perpendicular	$a \neq b \neq c$, 90, 90, 90	$P222_{(1)}$, $P222_1$, $P2_12_12$, $P2_12_12_1$; $C222_{(1)}$; F222; $I2_{(1)}2_{(1)}2_{(1)}$
tetragonal	P, I	one 4-fold axis ‖ ***c***	$a = b \neq c$, 90, 90, 90	$P4_{(1,3)}$, $P4_{(2)}$, $P4_{(2)}22$, $P4_{(2)}2_{(1)}2$, $P4_{(1,3)}22$, $P4_{(1,3)}2_12$; $I4_{(1)}$, $I4_{(1)}22$
trigonal	P	one 3-fold axis ‖ ***c***	$a = b \neq c$, 90, 90, 120	P3, $P3_{(1,2)}$, $P3_{(1,2)}21$, $P3_{(1,2)}12$
hexagonal	P	one 6-fold axis ‖ ***c***	$a = b \neq c$, 90, 90, 120	$P6_{(1,3)}$, $P6_{(1,5)}$, $P6_{(2,4)}$, $P6_{(1,3)}22$, $P6_522$, $P6_{(2,4)}22$
rhombohedral	R	one 3-fold axis	$a = b = c$, $\alpha = \beta = \gamma \neq 90$	R3, R32
cubic	P, F, I	four 3-folds along body diagonals	$a = b = c$, 90, 90, 90	$P2_{(1)}3$, $P4_{(1,3)}32$, $P4_{(2)}32$, $I2_{(1)}3$, $I4_{(1)}32$; F23, $F4_{(1)}32$

The "≠" sign means "not necessarily". Non-subscripted numbers signify *n*-fold rotation axes. Subscripts denote screw axes. For example, $P3_{(1,2)}$ is short for the two space groups $P3_1$ and $P3_2$.

In contrast to the space group symmetries, which are limited to a certain set of rotations and translations, the asymmetric unit of a crystal can be anything from a single atom to a huge complex, and may display any kind of symmetry. The asymmetric unit can be a truly asymmetric macromolecule, it can be an oligomer of *n*-fold symmetry, or it can be a subset of a symmetric oligomer (Figure 24.10). The term *asymmetric unit* therefore does not mean that it must be asymmetric, but that if it displays symmetry, this *local* or *non-crystallographic symmetry (NCS)* is restricted to the asymmetric unit and different from the global crystallographic symmetry of the space group. NCS is observed in about ⅓ of all macromolecular crystals. For example, the asymmetric unit of the GroEL structure contains two seven-membered rings. The NCS is 7-fold, something that is impossible for crystallographic symmetry (Figure 24.10). On the other hand, the asymmetric unit may also contain only a part of a symmetric oligomer if a symmetry axis of the oligomer coincides with a space group axis. An example is HIV–1 capsid protein, where the biologically relevant unit is a hexamer but the asymmetric unit contains only a monomer. This protein crystallizes in space group P6, and the hexamer is generated by the crystallographic 6-fold symmetry. (Figure 24.10). Thus, the content of the asymmetric unit has no relation to the biologically active particle, and it is sometimes difficult to decipher the biological unit from the crystal structure. Methods analyzing the molecules in solution are then required.

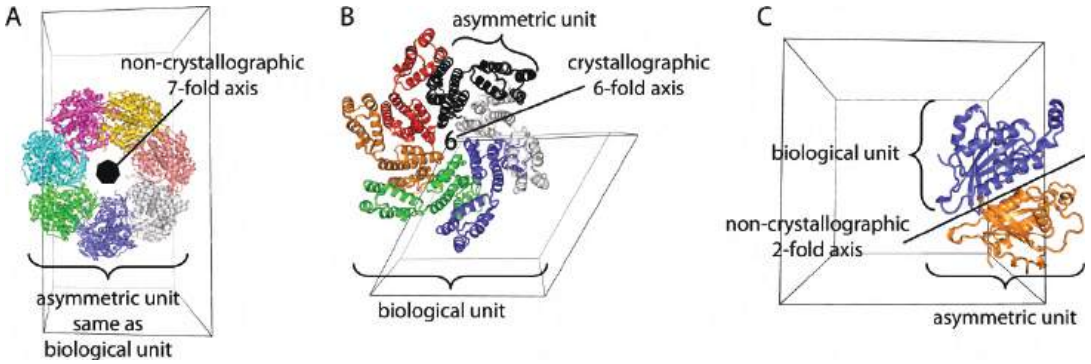


FIGURE 24.10: Asymmetric unit and biological unit. A: In the GroEL structure (PDB-ID 1sx3), the homo-heptamer is both the asymmetric and the biological unit. B: A single monomer is present in the asymmetric unit of the HIV-1 capsid structure in space group P6 (PDB-ID 4xfx). The biologically relevant hexamer is generated by the 6-fold crystallographic axis of this space group. C: The asymmetric unit may also contain more than one biological unit, which can be related by NCS. While the small GTPase Cdc42 forms a dimer in the crystal, it is a monomer in solution (PDB-ID 1a4r).

The first probable space group and cell dimensions are determined from the diffraction data (Section 24.4.3). Assuming the cell dimensions and the space group symmetry are correct, the likely content of the asymmetric unit can be estimated, *i.e.* how many molecules there might be. If this number is less than unity, the macromolecule does not fit into the asymmetric unit and the space group may be wrong. If there are several molecules in the asymmetric unit, this information can be used for electron density improvement by NCS averaging (Section 24.4.5). The number of molecules is estimated by calculating the *Matthews parameter* V_M, a reciprocal crystal packing density given in units of Å^3/Da (in crystallography, the Ångstrøm is still the preferred unit for length; it equals to 0.1 nm). V_M is calculated by dividing the volume of the asymmetric unit by the molecular mass M_M of the crystallized macromolecule. The volume of the asymmetric unit is the volume of the unit cell, V_{cell}, divided by the number of symmetry operators ($\#_{symops}$) of the space group.

$$V_M = \frac{V_{cell}}{\#_{symops} \cdot M_M} \qquad \text{eq. 24.2}$$

V_{cell} is calculated as the volume of a general parallelepiped:

$$V_{cell} = abc \cdot \sqrt{1 + 2\cos\alpha \cdot \cos\beta \cdot \cos\gamma - \cos^2\alpha - \cos^2\beta - \cos^2\gamma} \qquad \text{eq. 24.3}$$

In the case of all angles equal to 90°, eq. 24.3 reduces to $V_{cell} = a{\cdot}b{\cdot}c$. The molecular mass M_M is chosen as integer multiples of the crystallized molecule to obtain candidate V_M values for one, two, three, *etc.* molecules per asymmetric unit. V_M has a range of 2.2–4 Å^3/Da for most macromolecular crystals.

In summary, the entire macroscopic crystal is fully described by the asymmetric unit and the space group. This means that the crystal structure is determined once the structure of the asymmetric unit is known. All structure determination efforts aim at elucidation of the content of the asymmetric unit.

24.4 X-RAY CRYSTALLOGRAPHY

X-ray crystallography is by far the most successful structure determination method: > 90% of all macromolecular structures are determined with this method. In contrast to other techniques that deliver atomic models (NMR, electron microscopy), X-ray crystallographic objects can be of any size, from single atoms to many MDa objects such as ribosomes (see Figure 19.39) and viruses (see Figure 18.39). The quality and resolution of X-ray crystal structures (or short: crystal structures, if no confusion with electron or neutron crystallography can arise) is often high enough to infer enzymatic mechanisms, to reveal binding interfaces in large macromolecular complexes and to speed up structure-based drug design in search for new medications. The recent advent of high-resolution cryo-EM structures and protein models generated by machine learning does not (yet) change the prominent position X-ray crystallography has in biology. Cryo-EM is still too time-consuming (and expensive) for routine structure determination. Likewise, current algorithms for structure prediction (Section 20.3.2) currently are not accurate enough to answer biochemical questions that depend on positional errors of side chains < 0.05 nm. The three aspects can, however, be favorably combined to overall accelerate a project, for example by using a predicted model for molecular replacement phasing (Section 24.4.4.3) or stitching together of a multi-protein complex in a low-resolution cryo-EM map using high-resolution crystal structures of individual subunits.

X-ray crystal structure determination is a multi-step process involving sample preparation, crystallization, data collection, phasing, model building into electron density, followed by refinement, validation, and interpretation of the model (Figure 24.11).

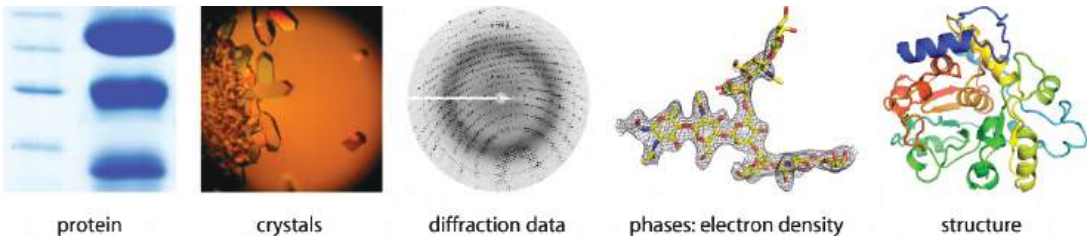


FIGURE 24.11: Steps in X-ray crystal structure determination. A macromolecule or a complex is purified at high concentration and crystallized. Diffraction data are collected, but during the process the phases of the diffracted light beams are lost. Phases are required to calculate an electron density map. The map is interpreted in terms of an atomic model. The model is often depicted as a ribbon diagram.

Since crystallization is a general requirement for all crystallographic techniques, it has been covered at the outset of this Chapter. Many aspects of data collection, phasing, model building, refinement, validation, and interpretation are the same for all crystallographic facets, so their discussion in this Section is also valid for electron and neutron crystallography. Differences specific for a technique, for example the generation of the electron and neutron beams and data collection specifics, are mentioned in the respective Section (Sections 24.7 and 24.8). Here, we start by

describing how X-rays are generated, then follow the general outline stated above for the remainder of the Chapter.

21.4.1 Generation of X-rays

There are two widely used methods to produce high-brilliance X-rays: (1) emission during electron transition from a high-energy state to a low-energy state between the inner shells of atoms, and (2) emission from particles accelerated to relativistic velocities. X-rays in the laboratory are generated using the first method in *sealed tubes* and *rotating anodes*. Orders of magnitude higher brilliances are obtained in *synchrotrons* that use the second method.

The intensity and quality of an X-ray beam is described by several parameters, among them the number of photons in the beam, the beam angular divergence, the cross-section of the beam, and its monochromaticity. These parameters are important for crystallography. The higher the intensity the smaller the data acquisition time. A small cross-section limits background scattering and allows the study of tiny samples. Small angular divergence improves the signal strength from the sample, and a precisely defined wavelength is essential for several structure determination methods.

The *flux* of an X-ray beam is the number of photons per second per 0.1% bandwidth, where 0.1% bandwidth at 0.1 nm would include photons of wavelengths between 0.0999 nm and 0.1001 nm. The *brightness* of a beam is the flux per $mrad^2$. The angular increment in millirad (mrad) is a measure of how quickly the beam diverges with distance from the source. Highly collimated beams where the photons are almost parallel to each other have high brightness. Finally, the *brilliance* is the brightness per mm^2, where mm^2 refers to the cross-sectional area of the source.

$$brilliance = \frac{\#\ photons}{s \cdot mm^2 \cdot mrad^2 \cdot 0.1\%\ bandwidth}. \qquad \text{eq. 24.4}$$

In other words, brilliance is the photon flux passing through 1 mm^2 area with a divergence given in millirad. High brilliance X-ray beams can be shaped with respect to cross-section, angular divergence, and monochromaticity while maintaining a high enough intensity for the experiment.

When Röntgen discovered X-rays in 1895, they were generated using the first method by electrons passing through a partially evacuated glass tube. The deflection, slowing, or sudden stopping of electrons by collisions with gas molecules in the glass tube produced a continuous X-ray spectrum, termed the *Bremsstrahlung* or *braking radiation*. Most applications require X-rays of defined wavelength, which are produced in the laboratory by shooting electrons from an *electron gun* onto a metal target (Figure 24.12). The electrons are evaporated from a cathode, a coiled metal filament heated to red-hot temperatures by an electric current of ~4 A. A magnetic field of 20–60 kV accelerates these primary electrons toward the metal-coated anode, where they eject other electrons from the inner K-shell of the anode atoms (Figure 24.12). The resulting "hole" is filled by an electron from an outer shell of higher energy, usually from the next higher energy L-shell. The energy difference between the electronic states is emitted as an X-ray photon. Each element has its *characteristic X-ray line spectrum* that superimposes on the continuous Bremsstrahlung (Figure 24.12). The choice of anode material determines the wavelength of the X-rays generated.

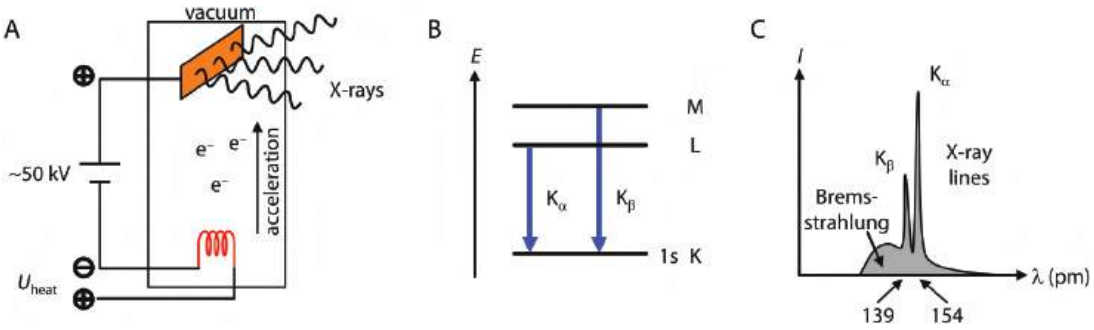


FIGURE 24.12: X-ray emission by intra-atomic electron transitions. A: Schematic representation of how electrons are evaporated from a cathode and accelerated toward the anode target in a sealed tube or rotating anode. The cathode, often a tungsten wire, is heated to 2000–3000°C by an electric current (red coil). The high-voltage electric field accelerates the electrons toward the anode. B: Energy states in an anode atom. The transition for lighter metals (Table 24.2) is usually between K-, L-, and M-shells. The emitted X-ray photon is termed α- or β-radiation when emitted from the L- or M-shell, respectively. C: The X-ray spectrum resulting from the impact of cathode electrons on the anode is a superposition of the continuous spectrum from scattering and the line spectra from ionization of the anode material (here copper).

The first X-ray sources, used in Röntgen's time, were *sealed tubes*, evacuated glass tubes that hold the anode and cathode. Higher X-ray intensities, used from the 1960s until the present, were achieved by increasing the electron flux onto the anode. The brilliance of the beams was improved by focusing the electrons on a smaller spot on the anode. As most of the energy of the impacting electrons is converted into heat, increase of the electron flux requires active cooling of the anode or else the impacting electrons will melt a hole in the anode. As the name implies, in a *rotating anode* generator, the anode is a rotating wheel, so the electrons hit the anode always at a different spot. A water circuit inside the anode ensures additional cooling of the anode material, but is a considerable source of machine failure. Due to the many moving parts in rotating anodes many laboratories now choose sealed tubes. Micro-focus sealed tubes achieve X-ray brilliances ($2{\cdot}10^9$ photons$\cdot$s$^{-1}\cdot$mm$^{-2}\cdot$mrad^{-2}) comparable to those of micro-focus rotating anodes ($6{\cdot}10^9$–$26{\cdot}10^9$ photons$\cdot$s$^{-1}\cdot$mm$^{-2}\cdot$mrad^{-2}) and require less maintenance. The X-rays emitted from the anode are focused by total reflection at confocal multi-layered mirrors.

TABLE 24.2:
X-ray K-lines of selected elements.

target	K_α (pm)	K_β (pm)
Cu	154	139
Cr	229	209
Fe	194	176
Co	179	162
Ni	166	150
Mo	63	71

The second widely used method to generate X-rays is by light emission from charged particles accelerated to relativistic velocities. This principle is the basis for the generation of X-rays in synchrotrons. Synchrotrons are large circular facilities, some with a diameter of several hundred meters, which produce highly focused light over a wide spectral range from infrared microwaves to hard X-rays. The X-rays produced by synchrotrons reach brilliances > 10^{21} photons s^{-1} mm^{-2} mrad^{-2}. Light emission in synchrotrons is a relativistic effect. If a charged particle traveling at almost the speed of light is further accelerated, it emits part of its energy as light. Forcing a charged particle

on an orbit by a magnetic field means continuous centripetal acceleration toward the center of the orbit, which results in continuous light emission. From each point on the orbit, the light is emitted tangentially as a tightly focused cone (Figure 24.13). The light is linearly polarized and has a continuous (white) spectrum.

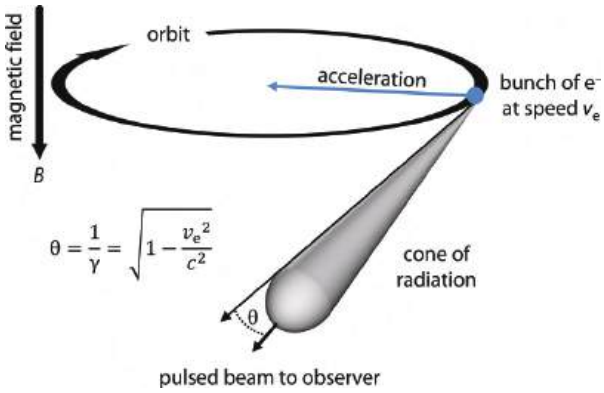


FIGURE 24.13: Light emission at relativistic speed. Charged particles traveling on a circular path experience continuous centripetal acceleration (arrow). Bunches of electrons orbit in synchrotrons at velocity v_e approaching the speed of light c. Energy is radiated in a focused cone tangentially to the orbit at all points where electrons are. The angle θ (in rad) of the cone depends on the velocity of the particle and is the inverse of γ. The light seen by the observer is pulsed with a frequency corresponding to the orbiting frequency of the electrons.

An important characteristic of a synchrotron X-ray beam is the Lorentz factor γ, the ratio of the total particle energy E_0 in the storage ring and its rest mass energy $m_e c^2$.

$$\gamma = \frac{E_0}{m_e \cdot c^2} \qquad \text{eq. 24.5}$$

γ is the "correction factor" for the increased mass m_{rel} of an electron moving at relativistic speeds v_e:

$$m_{rel} = \frac{m_e}{\sqrt{1-\frac{v_e^2}{c^2}}} = m_e \cdot \gamma \qquad \text{eq. 24.6}$$

The reciprocal value of γ is also known as the *natural emission angle* of synchrotron radiation. This value is equivalent to half the opening angle of the radiation cone (Figure 24.13) and serves as a measure of how focused the light beam is. The larger γ, the higher collimated is the radiation. To an observer the synchrotron light appears pulsed with a frequency equivalent to the orbiting frequency of the electrons.

In synchrotrons, an electron bunch from an electron gun, also known as a LINAC (linear accelerator) is injected into a small booster synchrotron. Once the electrons have reached their final velocity, the bunch is injected into the larger *storage ring*, where it is kept on a stable orbit by the strong magnetic fields of dozens of superconducting *bending magnets* with field strengths of > 1 T each (Figure 24.14). The synchrotron light emitted by the electrons is used at dozens of beamlines constructed tangentially to the storage ring. Flux, brightness, and wavelength of the X-rays are adjusted by filters, optics, and diffraction gratings located between the storage ring and the end station. A typical X-ray wavelength range for a synchrotron is 0.05–0.4 nm (0.5–4 Å).

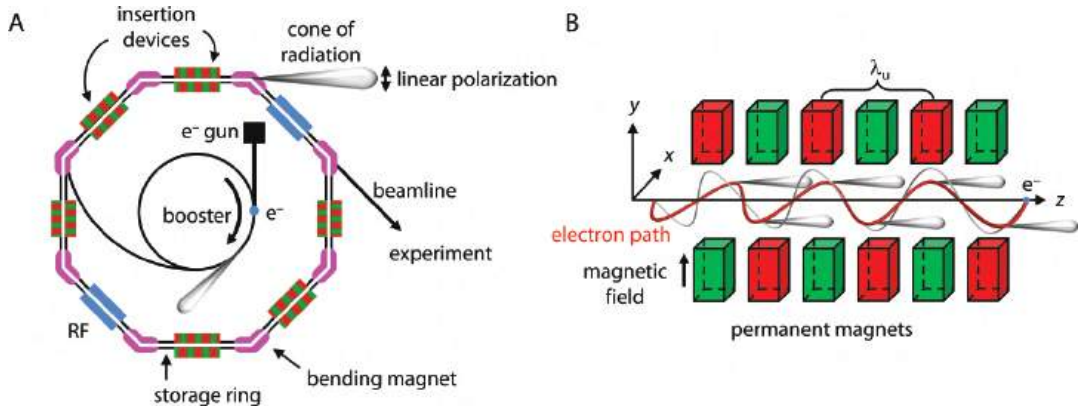


FIGURE 24.14: Synchrotron and insertion devices. A: Electrons from an electron gun are accelerated by a magnetic field and injected into a booster ring, where they are accelerated further by magnetic fields. The electrons are then injected into the storage ring where bending magnets (magenta) keep the electrons in orbit. The electron current of several hundred mA is usually maintained by "topping up" with electrons during operation. A continuous spectrum of linearly polarized synchrotron light is emitted in the forward direction, *i.e.* tangential to the orbiting electrons. Beamlines run tangential to the storage ring, and optics focus X-rays and select a certain wavelength for the experiment. Linear sections of the storage ring host insertion devices, termed *wigglers* and *undulators*, that deflect the electrons perpendicularly from their path in the plane of the storage ring, which intensifies the synchrotron light. The radiated energy that is lost by the electrons is replenished by radiofrequency coils (RF). B: Schematic view of wigglers and undulators. Magnets can be electromagnetic with current-carrying coils or permanent, made from *e.g.* samarium-cobalt or neodymium-iron-boron. Magnets are shown with north and south poles as red and green, spaced at a period of λ_u. The magnets construct an oscillating magnetic field (gray line) that sinusoidally deflects the orbiting electrons. As this deflection is equivalent to a net acceleration, the intensity of the emitted synchrotron light is boosted. The situation shown here is that in a wiggler, where the radiation cones do not overlap and the intensities are just added to produce non-coherent light of higher intensity. In an undulator, the deflection is not as strong, so the radiation cones do overlap. The polarized light emitted in an undulator at each peak is in phase with the other peaks. The resulting constructive interference further boosts the brilliance of the now coherent light. If the magnet arrays can slide with respect to each other, the deflection path for the electrons can be made helical. The result is circularly polarized synchrotron light with the helicity of the deflection path determining the helicity of the light.

A storage ring is not circular but a polygon with dozens of linear sections connected by the same number of short circular arcs. Bending magnets sitting at these junctions change the direction of the orbiting electrons from one linear segment to another. While the main task of a bending magnet is to keep the electrons on their closed-path orbit in the storage ring, the resulting burst of electromagnetic radiation induced by a bending magnet has a brilliance of about 10^{13} photons·s^{-1}·mm^{-2}·mrad^{-2} at 0.1% bandwidth. Even higher brilliances and much more focused light are produced by *insertion devices* known as *wigglers* and *undulators*, which increase the brilliance of the synchrotron light to about 10^{16} and 10^{19} photons·s^{-1}·mm^{-2}·mrad^{-2}, respectively. Upgrades in the 2020s of synchrotrons to their "third generation" further boost brilliances to 10^{21}–10^{24} photons·s^{-1}·mm^{-2}·mrad^{-2}. Insertion devices are placed (inserted) at linear sections of the synchrotron, where experimental hutches receive the lights for measurements. Wigglers and undulators are *multipole magnets*, *i.e.* linear arrays of magnets with alternating polarity (Figure 24.14) that deflect the electron beam of the storage ring passing through their alternating magnetic field. A deflection of $\gg 1/\gamma$ is characteristic of a wiggler, whereas an undulator deflects the electron beam with $\leq 1/\gamma$. Both devices convert kinetic energy of charged particles into light. The names come from the periodic deflection of the electron trajectory by the Lorentz force when they pass the device (they "wiggle" and "undulate"). Since the deflection amounts to a net acceleration of the electrons within the device, the electrons emit light of high intensity at each magnetic pole. In a wiggler, the total light intensity is therefore proportional to the number N of magnets (a few to more than 100). Compared to bending magnets, wigglers produce about 1000-fold more brilliant light. Undulators, on the other hand, achieve yet

another 10^3–10^6-fold increase in brilliance over wigglers because the light intensity emitted by them scales with N^2. The difference between a wiggler and an undulator lies in the interference of light emitted at the magnetic poles. While no strong interference is present in wigglers (they produce incoherent light due to the larger deflection of the electron beam), undulators are constructed such that light emitted at one magnetic pole constructively interferes with the light emitted from another magnetic pole. The light emitted by an electron at one magnetic pole travels faster than the electron. If the time difference between the electron and the emitted ray to reach the next magnetic pole corresponds to traveling one wavelength, constructive interference occurs. The interference condition for the light rays depends on the spacing of the magnets, λ_u, the energy of the electron beam, $\gamma m_e c^2$ (eq. 24.5), and the strength of the magnetic field. The magnetic field can be adjusted by changing the distance, or *undulator gap*, between north and south poles of the magnets. Due to the interference, undulators produce a coherent beam consisting of a fundamental frequency and several higher-frequency harmonics, one of which is chosen for the experiment by a monochromator. Undulators are now routinely used in synchrotron beamlines for SAXS and X-ray crystallography. The brilliance of light produced by undulators is ten orders of magnitude higher than cathode/anode devices, and when arranged in a series or enlarged to hundreds of meters, give rise to even more powerful *free electron lasers* (FELs; Section 24.6) with brilliances of about 10^{31}–10^{34} photons·s^{-1}·mm^{-2}·mrad^{-2}.

24.4.2 X-RAY DIFFRACTION FROM CRYSTALS

Because all molecules in a crystal are in defined relative orientations, the X-rays scattered from different molecules have a fixed phase relationship. This phase relationship leads to strong interference of the scattered beams with the result that crystals diffract X-rays only in certain directions. Geometrically, the diffraction of X-rays by crystals can be described as if the light were reflected at sets of equidistant parallel planes that run through the crystal and are oriented at a certain angle θ relative to the incident beam (Figure 24.15). Using this concept of reflection at planes, we can calculate under which conditions the reflected X-rays are in phase, interfere constructively, and can thus be measured as a *reflection* (a spot on a detector corresponding to a certain number of photons).

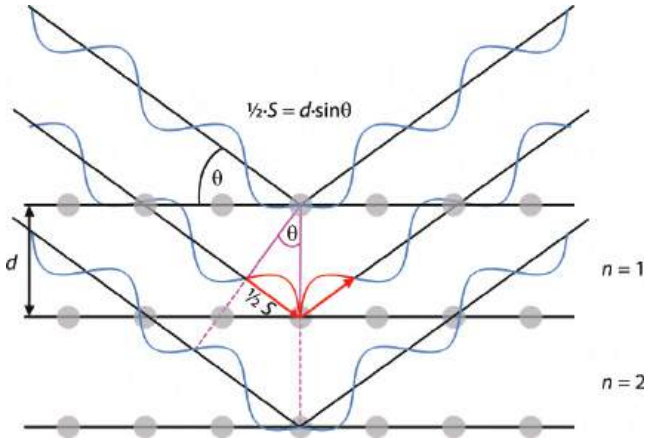


FIGURE 24.15: Bragg's law. Three planes of distance *d* are shown side-on, passing through atoms (gray dots). X-rays (blue waves) are reflected from these planes. The total reflection angle is 2θ. Constructive interference is only possible if the path difference *S* is an integer multiple of the wavelength: 1·λ between the first and second plane ($n = 1$), 2·λ between the first and third plane ($n = 2$), *etc.*

Reflection of X-rays at the first plane follows Snellius' law (Section 25.4; eq. 25.18): the angle of incidence is identical to the emergent angle (Figure 24.15). Thus, the total reflection angle is 2θ. X-rays reflected by the second plane travel a path difference *S* with respect to the first X-rays. In order for the two X-ray beams to remain in phase, *S* must be an integer number of the wavelength:

$S = n \cdot \lambda$. By trigonometry, we obtain $\frac{1}{2} \cdot S = d \cdot \sin\theta$, where d is the interplanar spacing. Combining these two expressions, we obtain Bragg's law:

$$n \cdot \lambda = 2 \cdot d \cdot \sin\theta \qquad \text{eq. 24.7}$$

Bragg's law states that positive interference at an angle θ is observed as a *reflection*, if the path difference of X-rays reflected from planes of spacing d is an integer multiple of the wavelength λ. In other words, only if the path difference of the X-rays is an integer number of wavelengths will the reflected beams interfere constructively. If the path difference of X-rays reflected from the second plane is but only slightly different from an integer multiple of the wavelength compared to the first plane, negative interference will occur. Because there are so many planes in a crystal, there will always be a plane "further down" in the crystal where the phase difference is exactly 180° and the reflected X-rays will cancel. While in solution, the light is scattered in all directions, it is diffracted only in certain directions by a crystal. In the absence of absorption, the total light intensity remains unchanged, so the few X-rays that interfere constructively are enormously enhanced compared to solution scattering, and the reflection intensities are therefore readily measurable.

The interplanar distance d corresponds to the resolution of the diffraction experiment: the smaller d, the higher the resolution. The resolution is limited to $\lambda/2$ by the fact that $\sin\theta_{max} = 1$ for $\theta = 90°$, in accordance with the limiting resolution in optical microscopy (see Section 25.1.3).

For $n = 1$, Braggs law states that each *set of planes* with interplanar distance d gives rise to a single reflection at angle θ. The set of planes is indexed by three integer numbers, termed the *Miller indices* h, k, and l (Figure 24.16). Many sets of equidistant planes can be drawn through a crystal lattice, but all planes within a set must intersect the unit cell axes rationally, or diffraction from different unit cells will lead to negative interference. The indices h, k, and l count the number of times that the set of planes cuts the unit cell axes. The same indices are also used to refer to the X-ray wave reflected from the set of planes, the reflection (hkl) with its intensity I_{hkl}. It is important to note that the larger the indices, the closer together lie the planes. Since closer planes mean more structural detail, reflections with larger indices have a higher resolution.

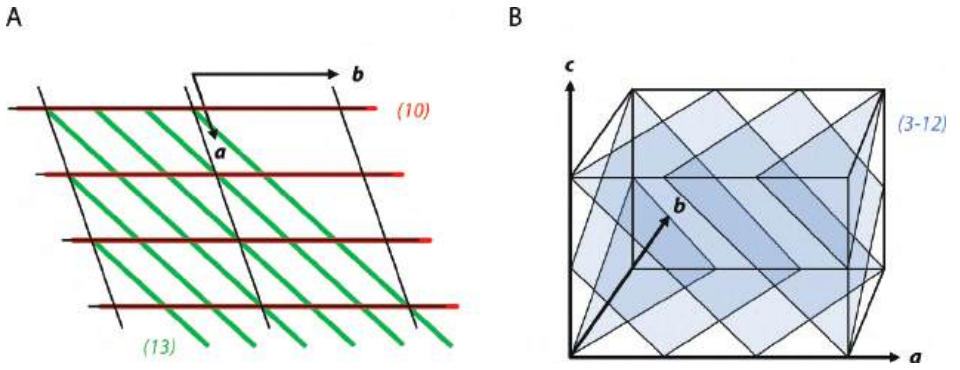


FIGURE 24.16: Lattice planes and Miller indices. A: A three-dimensional lattice (black lines) is projected along the *c*-axis so it appears two-dimensional for simplicity. Planes parallel to the *c*-axis then show as lines (green and red). The unit cell constants *a* and *b* are shown as arrows. The green and red planes are two of the many possible sets of lattice planes that rationally intersect the *a*- and *b*-axes. Successive planes of the green set intersect the lattice at spacings of *a* in the *a*-direction and at spacings of ⅓·*b* in the *b*-direction. Thus, there is one plane every distance *a*, and three planes every distance *b*, hence the set of planes is named (*13*) in two dimensions. Since the set of red planes is parallel to the *b*-axis, it will never intersect it, so the index along this direction is zero. The red set also has a spacing of *a* along the *a*-direction, so this set is indexed as (*10*). In three dimensions the green and red sets have indices (*130*) and (*100*). B: The three-dimensional example shows a (*3–12*) set of planes. The origin of the unit cell is at the bottom left. The negative index (*–1*) for the *b*-intersection comes from the fact that the planes intersect the *b*-axis in the negative direction.

Diffraction of X-rays into defined directions does not depend on the content of the unit cell, only on the symmetry of the crystal. However, the unit cell contents determine the reflection intensity I_{hkl}. In other words, the distribution of the reflections is determined by the space group and cell dimensions, and the reflection intensities are determined by the contents of the asymmetric unit. Mathematically, the observed diffraction pattern (see Figure 24.19) is a result of the convolution of the crystal lattice with the continuous molecular transform.

We mentioned before that the scattering process occurs in reciprocal space (Section 23.2.2), with the amplitude of the scattering vector in units of reciprocal length (nm^{-1}). The same real/reciprocal space relation holds for diffraction, and again the mathematical connection between the two spaces is the Fourier transformation. The set of planes with interplanar spacing *d* can be represented by a single vector that is perpendicular to the planes and has the magnitude *1/d*. Doubling, tripling, *etc.* of *d* will lead to the reciprocal magnitudes *1/(2d)*, *1/(3d)*, *etc.* The end points of the perpendicular vectors span a new lattice, known as the *reciprocal lattice* (Figure 24.17). The (*hkl*) triplets characterize equally the planes in real space and their associated planes in reciprocal space. The concept of a reciprocal lattice may be difficult to grasp, but it has the advantage that a whole set of planes in a crystal is described by a single reciprocal lattice point at which the reflection intensity is measured, and that it highlights the reciprocal relationship between the diffraction pattern and the real space crystal: a real space (direct) lattice with closely spaced points corresponding to small unit cell dimensions will give a diffraction pattern with widely spaced reflections corresponding to large reciprocal cell dimensions. It is a general property of the Fourier transformation to invert the relation between parameters.

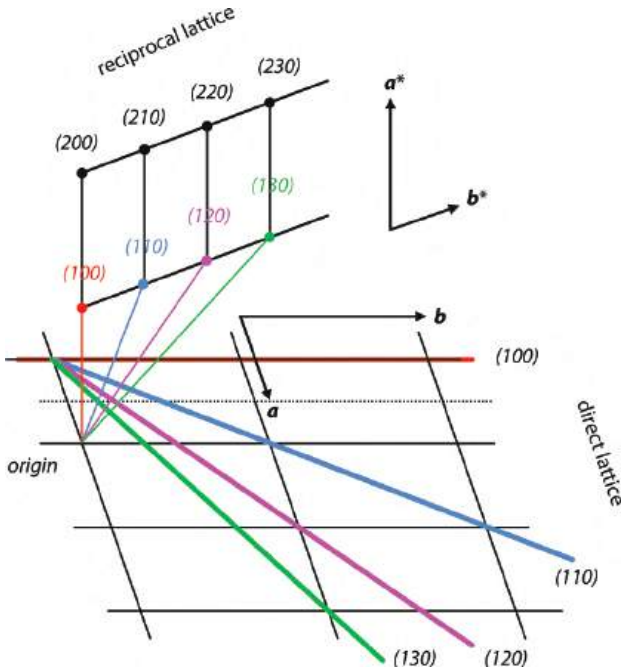


FIGURE 24.17: Real space and reciprocal space. A three-dimensional lattice is shown projected along the *c*-axis, so it appears two-dimensional. The colored lines are lattice planes parallel to the *c*-axis. Each line represents the whole set of planes (*hkl*). A vector perpendicular to the planes with magnitude *1/d* marks a reciprocal lattice point (a colored dot). The largest distance *d* characterizes the red set of planes, hence the reciprocal vector (*100*) is the shortest. Doubling the magnitude of the red vector (end point then at *200*) means halving *d* for the real space planes (the dotted line). Repeating this process for all possible sets of real space planes constructs the reciprocal lattice. The origin of the real and reciprocal lattices is the same, and the real space unit cell dimensions (*a*, *b*) are reciprocal to the reciprocal unit cell dimensions (*a**, *b**).

In solution scattering, the molecular form factor describes the combined scattering from all atoms within a molecule. Due to the ordering of molecules in crystals, the scattering by crystals is influenced not only by interference of waves within one molecule, but also by interference of scattered waves from different molecules. The analogy to the form factor in SAS is the *structure factor* in crystallography, which is given by

$$\boldsymbol{F}_{\mathbf{hkl}} = \left|F_{\mathrm{hkl}}\right| \cdot e^{i\phi_{\mathrm{hkl}}} = \sum_{n=1}^{N} f_{\mathrm{n},\theta_{\mathrm{hkl}}} \cdot e^{2\pi i \cdot (hx_{\mathrm{n}} + ky_{\mathrm{n}} + lz_{\mathrm{n}})} \qquad \text{eq. 24.8}$$

$\boldsymbol{F}_{\mathbf{hkl}}$ is the diffracted beam of all the N atoms in the unit cell in the direction given by the plane of indices hkl. Each atom n contributes to the scattering according to its resolution-dependent atomic scattering factor $f_{\mathrm{n},\theta\mathrm{hkl}}$. We recall from Section 23.2.1 that the magnitude of the scattering factor depends on the scattering angle θ. For diffraction, this means that the structure factor depends on the angle θ_{hkl} of the incident X-rays to the plane characterized by indices hkl (see Figure 24.16), as we have seen with Bragg's law. An important notion from eq. 24.8 is that the phase of the structure factor is only determined by the atomic positions $(x_{\mathrm{n}},y_{\mathrm{n}},z_{\mathrm{n}})$, where x_{n}, y_{n}, and z_{n} are the fractional coordinates of the n^{th} atom in the unit cell. If we know the atomic positions $(x_{\mathrm{n}},y_{\mathrm{n}},z_{\mathrm{n}})$, we can calculate the structure factor. However, in crystallography, the problem is the reverse: we would like to know the atomic positions based on the knowledge of the structure factor. Unfortunately, measured diffraction data contain only the intensities I_{hkl}, which are related to the square of the structure factor amplitude $|F_{\mathrm{hkl}}|$, but they do not contain the relative phase information ϕ_{hkl} for the reflections. This phase problem has to be solved in order to calculate the electron distribution in the unit cell of the crystal, which then allows us to place atoms $(x_{\mathrm{n}},y_{\mathrm{n}},z_{\mathrm{n}})$.

Until now, we have not used the complex notation for an electromagnetic wave because we were only dealing with amplitudes (proportional to intensities), not the phases. In crystallography, the phases are central to calculating the electron density, so we need to use the general formulation for electromagnetic waves. The structure factor $\boldsymbol{F}_{\mathbf{hkl}}$ is an electromagnetic wave characterized by its amplitude $|F_{\mathrm{hkl}}|$ and its phase ϕ_{hkl}, which can be visualized in an Argand diagram (Section 31.11; Figure 24.18).

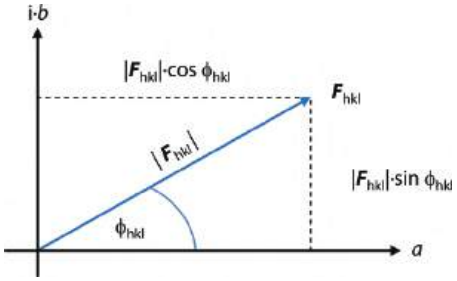


FIGURE 24.18: Structure factor representation. The two properties amplitude and phase of the structure factor are conveniently represented in an Argand diagram.

The atomic scattering factor (see Figure 23.7) is the Fourier transform of the electron density of a spherical atom. Since X-rays are scattered by electrons, the total diffraction of X-rays by a crystal is the sum of the contributions by all electrons in the unit cell, and consequently the Fourier transform of the total diffraction returns the electron density distribution of the unit cell. Reformulation of eq. 24.8 replaces the atomic scattering factors $f_{n,\theta hkl}$ with the electron density ρ_{xyz} at a point (x,y,z).

$$\boldsymbol{F}_{\mathbf{hkl}} = \left|F_{\mathrm{hkl}}\right| \cdot e^{i\phi_{\mathrm{hkl}}} = V \cdot \sum_{\mathrm{xyz}} \rho_{\mathrm{xyz}} \cdot e^{2\pi i \cdot (hx+ky+lz)} \qquad \text{eq. 24.9}$$

The electron density in the unit cell is then the Fourier transform of the structure factor:

$$\rho_{\mathrm{xyz}} = \frac{1}{V} \cdot \sum_{\mathrm{hkl}} \left|F_{\mathrm{hkl}}\right| \cdot e^{i\phi_{\mathrm{hkl}}} \cdot e^{-2\pi i \cdot (hx+ky+lz)} \qquad \text{eq. 24.10}$$

In contrast to the structure factor, which is a complex number, the electron density is a real number. If we evaluate eq. 24.10 at every point (x,y,z) in the unit cell, we have defined the electron distribution of the entire crystal (by simple translation). In practice, it is sufficient to know the electron density distribution (representing the atomic structure) of the asymmetric unit, because the unit cell is built from the asymmetric unit by the symmetry operators of the space group. There are two phase factors in eq. 24.10, one is the phase ϕ_{hkl} for the diffracted X-ray, the other is a standing wave (hx+ky+lz) repeating each unit cell (2π) with a wavelength corresponding to integer fractions of the unit cell dimensions. It is the first phase, ϕ_{hkl}, that is missing from the measured data. Sections 24.4.3 and 24.4.4 will discuss what information can be gleaned from the intensity data without phases before continuing to discuss methods to retrieve the phases.

24.4.3 Diffraction data collection and analysis

The general setup for diffraction data collection is the same as that for SAS (Figure 23.8): the sample is placed in the X-ray beam and the scattered/diffracted light intensity is measured as a function of θ. While molecules in solution have random orientations, their relative orientations are fixed in crystals. Therefore, the difference to solution scattering is that the crystal has to be rotated in order to collect all the diffracted light. For this purpose, the crystal is mounted on a *goniometer* (Figure 24.19) that enables rotation of the crystal about one or more axes relative to the incident X-ray beam. For room temperature data collection (rarely done nowadays), the crystal is transferred into an X-ray transparent capillary and padded with mother liquor to avoid drying out. For the standard cryogenic data collection (Box 24.2), crystals in cryo-protectant are scooped up in a plastic loop at the end of a magnetic pin. The loop is rapidly cooled in liquid N_2 and placed on a magnetic base of the goniometer (Figure 24.19). The crystal is kept at 100 K in a cryo-stream of nitrogen gas and centered in the X-ray beam. A few hundred diffraction images are collected, which constitute a *dataset*. During each image collection, the crystal is rotated by a small angular increment of 0.05–1°, termed the *oscillation range*. The total oscillation range required for a dataset depends on the symmetry of the space group. The more symmetric the space group, the less angular space has to be covered for a dataset. For a highly symmetric cubic lattice, about 22° total rotation can be sufficient, but for the least symmetric triclinic lattice at least 180° are necessary.

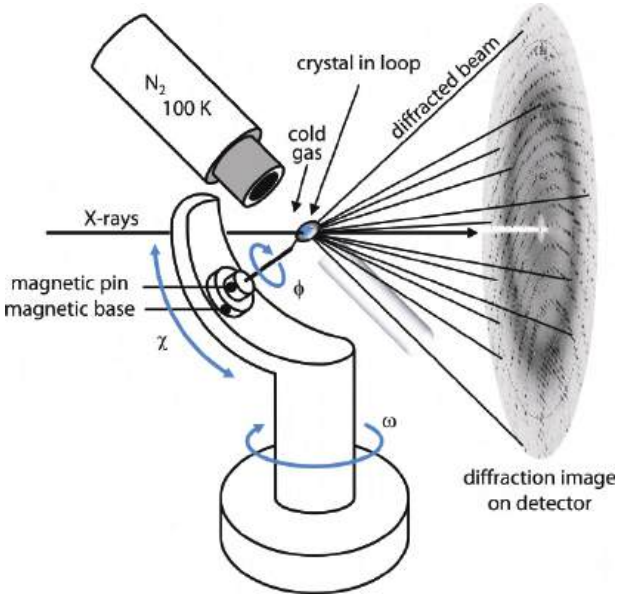


FIGURE 24.19: Diffractometer and cryo-cooling. The crystal is in a loop at the end of a magnetic pin that fits onto the goniometer. A laminar N_2 gas stream keeps the crystal at 100 K for the duration of the experiment. While some goniometers have only a single axis of rotation (ω), three-circle goniometers can rotate the crystal about three axes, termed ω, ϕ, and χ. The diffraction image on the right-hand side is not to scale.

Apart from the content of the crystal, the intensity of the reflections, I_{hkl}, increases with the order of the crystal, the intensity of the primary beam, the exposure time, and the time the crystal is kept in reflection condition. Bragg's law states that the reflection condition for an interplanar spacing is fulfilled at one exact angle θ. When continuously rotating a perfectly ordered crystal, this condition would only be fulfilled for an infinitely small time, and the intensity collected would be zero. However, macroscopic crystals are imperfect: they are a mosaic of perfect micro-crystals that are rotated slightly with respect to each other in all three dimensions. The mean angular rotation of the mosaic regions is termed the *mosaicity*, which is responsible for the fact that the crystal remains in reflection condition for about 0.1–2° angular rotation. This is large enough to collect the reflection intensities during continuous rotation with oscillation speeds between 0.01° s^{-1} and 1° s^{-1}. A typical diffraction image on a CCD (charge-coupled device; see below) detector is collected with an oscillation range of 0.5° and exposure times of 0.5–2 s. In contrast, detection with modern pixel detectors enables a complete dataset to be collected in thin slices of typically 0.05–0.25° per image with an angular velocity of 1° s^{-1} in a matter of minutes (3 min for 180°).

The first X-ray detectors used photographic film, which required tedious development and digitization of the reflection intensities. *Image plates* store the X-ray photon energy in pixels of a *phosphor* (crystalline BaFBr doped with a lanthanide) that luminesces after excitation with a suitable He-Ne laser. The phosphorescent light is amplified by a photomultiplier and counted. The detector is then erased by irradiation at high intensity to prepare it for the next exposure. The time for a readout/erasure cycle is about 1 min, which makes image plate detectors too slow for data collection at synchrotrons, but they are still used in some X-ray laboratories. CCDs follow an analogous strategy to image plates in that they store the X-ray photon energy in a phosphor: an X-ray photon liberates a fixed number of electrons in the respective pixel, which are attracted toward an anode, with the result of stable charge separation until readout. Readout of the CCD is initiated by alternating the polarity of a series of electrodes across detector lines (or columns, depending on the setup). The electron clouds travel along the lines, and at the end this current is converted into a voltage and recorded. CCD readouts are much faster (ca. 1 s) than those from image plates. The area covered by CCDs is smaller than that of image plates. If larger detector areas are required, several CCDs can be assembled into a 2x2 or 3x3 grid. Until the mid-2000s, such composite CCD detectors were the standard at synchrotron beamlines. The latest developments are single photon counting, or *direct pixel detectors*, which directly absorb and count X-rays in a silicon diode layer. Since every diode is connected to its own readout electronics (further discussed in Section 26.5), the readout dead time is on the order of a few μs and whole datasets can be collected in seconds. Pixel detectors have a dynamic range more than ten times that of CCDs, no electronic noise, and can be operated in continuous readout mode without closing the shutter between successive images, which facilitates time-resolved diffraction experiments (Section 24.5). These detectors are now the standard at synchrotrons.

Once the first few images have been collected, the positions of the reflections in the diffraction patterns can be *indexed*, *i.e.* ascribed their (*hkl*) values using the best fitting Bravais lattice, crystal system, and unit cell constants. At this stage, it is decided whether the lattice is likely primitive, C-centered, *etc.* and whether the crystal system might be triclinic, monoclinic, *etc.* (Table 24.1). Using this information, a *data collection strategy* can be calculated that tells how to rotate the crystal with respect to the incident X-rays in order to rapidly collect a complete dataset with minimal X-ray exposure of the crystal for minimal radiation damage. Triclinic crystals of the lowest symmetry require data collection over a range of 180° (360° if anomalous data is required), while highly symmetric cubic crystals can deliver a complete dataset from only 45° oscillation range or less. In general, the larger the number of symmetry operators of a space group, the smaller the oscillation range required for a complete dataset. Radiation damage is therefore less of a problem for high-symmetry space groups than for space groups of low symmetry.

Data processing involves integration of the reflections on all diffraction images and merging them into a list of intensity measurements I_{hkl} with their associated errors $\sigma(I_{hkl})$. The errors are the mean standard deviations for *symmetry-related reflections*, which are reflections that come from

different but crystallographically equivalent planes and therefore should have identical intensities. Symmetry-related reflections, or *symmetry-mates* for short, are almost always measured more than once. The average number of times a reflection or its symmetry-mate has been measured is the *multiplicity n* (formerly "redundancy"), with larger *n* giving more reliable mean intensities. Averaging (merging) the symmetry mates yields the *unique reflections* with their mean intensities $\langle I_{hkl} \rangle$ and standard deviations $\sigma\langle I_{hkl} \rangle$. The overall quality of the unmerged collected diffraction data may be expressed in the form of residual (*R*-) factors and as correlation coefficients. *R*-factors should be small (< 0.1) and correlation coefficients should approach unity. One of the most frequently used *R*-values is R_{sym}:

$$R_{sym} = \frac{\sum_{hkl} \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_{j=1}^{n} I_{hkl,j}} \qquad \text{eq. 24.11}$$

Here, a reflection from the plane (*hkl*) of intensity I_{hkl} is measured *j* times, but at most *n* times. The *j* reflections include the same reflection measured several times and all symmetry-mates. In practice, their intensities differ due to measurement errors and radiation damage. Their mean intensity $\langle I_{hkl} \rangle$ is subtracted from each individual intensity $I_{hkl,j}$, and the absolute value of the difference is taken. Summation over all unique reflections and normalization yields R_{sym} (eq. 24.11). R_{sym} has the unfortunate property of increasing with increasing multiplicity *n*, but for the overall data quality a large multiplicity is beneficial. An *R*-factor that takes the multiplicity into account is R_{meas}:

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \cdot \sum_{j=1}^{n} \left| I_{hkl,j} - \langle I_{hkl} \rangle \right|}{\sum_{hkl} \sum_{j=1}^{n} I_{hkl,j}} \qquad \text{eq. 24.12}$$

R-values are often calculated in resolution shells to monitor the deterioration of the data toward higher resolution. At low resolutions up to 1 nm, where reflections are strong and accurately measured, R_{sym} and R_{meas} should not exceed 10% (ideally 3–5%) or the choice of space group for the data processing might be wrong. With increasing resolution, *R*-factors also increase because the intensities I_{hkl} decrease and the differences in the numerator of eq. 24.11 and eq. 24.12 become large. *R*-factors are a good way to judge low-resolution data and R_{meas} can even be used to compare data from different crystals, but the practical high-resolution limit of diffraction data is better judged by using correlation coefficients.

Correlation coefficients test the consistency of two arrays of data. Diffraction data are analyzed for internal consistency by calculating the half-dataset correlation coefficient $CC_{1/2}$. The calculation may be done for the entire dataset or in resolution shells to follow the (expected) decrease of $CC_{1/2}$ as a function of resolution. For $CC_{1/2}$ calculation, the unmerged intensity data is split into halves 1 and 2, and the correlation between the two sets calculates to:

$$CC_{1/2} = \frac{\sigma^2_{\langle I_{hkl,1/2} \rangle} - 0.5 \cdot \overline{\sigma^2_{\langle I_{hkl,1/2} \rangle}}}{\sigma^2_{\langle I_{hkl,1/2} \rangle} + 0.5 \cdot \overline{\sigma^2_{\langle I_{hkl,1/2} \rangle}}} \qquad \text{eq. 24.13}$$

Here, $\sigma^2_{\langle I_{hkl,1/2} \rangle}$ is the variance of all the averaged intensities and $\overline{\sigma^2_{\langle I_{hkl,1/2} \rangle}}$ is the average of the variances of the averaged intensities. Several ways of halving the data can be envisaged. Splitting at the midpoint during data collection allows assessment of radiation damage, which increases over time. $CC_{1/2}$ will be small if radiation damage is present and has reduced the intensities in the data collected later in time (half 2). More often interleaving of the data (even *versus* uneven image numbers, for example) or just random splitting is used to divide the intensities and $CC_{1/2}$ now reports on internal consistency throughout the data, independent of time. Values for $CC_{1/2}$ should be > 0.95, but even $CC_{1/2} > 0.3$ is deemed statistically significant and is often used as a criterion for the practical resolution limit. If the diffraction data are cut at this $CC_{1/2}$, the signal-to-noise ratio $I_{hkl}/\sigma(I_{hkl})$ of reflections

in the highest resolution shell usually falls between 0.5 and 2.0. $CC_{1/2}$ is based on unmerged intensities. A related coefficient CC^* that reports on the correlation of the merged intensities *versus* the (unknown) true intensities is defined as:

$$CC^* = \sqrt{\frac{2 \cdot CC_{1/2}}{1 + CC_{1/2}}} \qquad \text{eq. 24.14}$$

Assuming CC^* = 0.5, $CC_{1/2}$ calculates to 1/7, or 0.143. This value for $CC_{1/2}$ is also known as the *Fourier Shell Correlation* (FSC) cut-off, and is used as a popular high-resolution criterion in cryo-EM (Section 26.5.3).

Once the diffraction data have been integrated in the correct space group, a very useful piece of information can be calculated using the measured intensities alone: the *Patterson function*. We recall that the electron density is the Fourier transform of the structure factor:

$$\rho_{xyz} = \frac{1}{V} \cdot \sum_{hkl} |F_{hkl}| \cdot e^{i\phi_{hkl}} \cdot e^{-2\pi i \cdot (hx+ky+lz)} \qquad \text{eq. 24.10}$$

In the absence of phases, we can calculate the Patterson function by using only the measured intensities $I_{hkl} \sim |F_{hkl}|^2$ as the *Fourier coefficients*:

$$P_{xyz} = \frac{1}{V} \cdot \sum_{hkl} |F_{hkl}|^2 \cdot e^{-2\pi i \cdot (hx+ky+lz)} \qquad \text{eq. 24.15}$$

Mathematically, the Patterson function is the convolution of the electron density with itself. This convolution is positive at points that correspond to interatomic vectors and zero elsewhere. Thus, the Patterson function corresponds to a set of vectors between atoms in a given volume V (a single molecule, an asymmetric unit, or a unit cell). A plot of this function, the *native Patterson map* (Figure 24.20), is a vector map where each vector magnitude corresponds to the distance between two atoms. The end points of the vectors tell about the relative positions of the atoms. The peak heights of the Patterson function are proportional to the product of the number of electrons of the atoms involved. As a consequence, very heavy atoms are more easily identified in the map than light atoms (see heavy atom derivatives; Section 24.4.4.1).

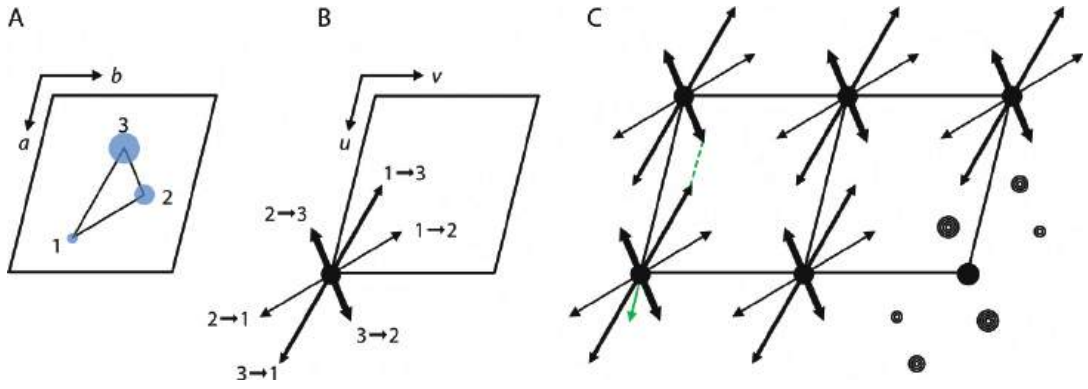


FIGURE 24.20: Patterson map. A: A hypothetical triangular molecule of three atoms with different mass is shown as blue dots of different radii in a two-dimensional unit cell. Atom 3 is the heaviest atom (highest electron density). B: Interatomic vectors are generated and translated to the origin. The unit cell has the same dimensions, but since the vectors denote relative, not absolute, coordinates, the coordinates of their end points are denoted by (*u,v,w*) rather than (*x,y,z*). Since the vector 1→2 has the opposite direction to 2→1, the Patterson map has a center of inversion symmetry at its origin (black dot): it is centrosymmetric. The peak height at the tip of the vector between atoms 2 and 3 is the largest off-origin peak, indicated by the thickness of the arrow. C: Translation of the unit cell to generate the lattice shows that there are also vectors between atoms located in adjoining unit cells (one example is drawn in green). The vectors in the lower right-hand unit cell are shown as contour lines, an equivalent way of plotting the Patterson map.

The procedure to draw a Patterson map starting from a given molecular structure is conceptually simple (Figure 24.20). Each vector between pairs of atoms is moved such that it starts at the origin. For N atoms, this will result in N^2 vectors, arranged centrosymmetrically about the origin. The largest peak is at the origin, because it collects all the N *self-vectors*, the vectors of zero length from one atom to itself. If the Patterson map is calculated over a larger volume than a single molecule, additional *intermolecular* vectors between atoms from adjacent unit cells are present in the map (Figure 24.20), which are important in molecular replacement phasing (Section 24.4.4.3). Although the Patterson map does not contain phase information, it does in principle contain all the information necessary to determine the crystal structure. Working back from a Patterson map to the structure is feasible for small molecules that have only a few atoms. With increasing number of atoms, the number of peaks in the Patterson map grows quickly, and they start to overlap, rendering assignment and direct structure determination impossible. We will see in Section 24.4.4 how the Patterson function is applied for phasing of macromolecular structures.

24.4.4 Phasing methods

Both intensities and phases are required for calculating the electron density. Several methods have been developed to retrieve the missing phases of the measured reflections. Three methods will be described, *isomorphous replacement*, *anomalous scattering*, and *molecular replacement*, all of which make use of the Patterson function. Isomorphous replacement and anomalous scattering simplify the problem of locating many atoms in the asymmetric unit by first locating only a small *substructure* of heavier atoms. The phases for the substructure then serve as starting phases for the macromolecule. In molecular replacement, the phases are calculated from a closely related structure that has been positioned in the unit cell of the unknown structure. It should be noted that molecular replacement currently is the most frequently used method to estimate phases, and with the advent of accurate protein model predictions (Section 20.3.2) will further increase in importance. It might be that the elegant traditional methods of isomorphous replacement and anomalous diffraction become historic soon.

24.4.4.1 Isomorphous replacement

The first step in isomorphous replacement is to obtain a heavy atom derivative of the crystal, where each macromolecule has bound one or just a few heavy atoms (Box 24.3). A dataset of the heavy atom derivative (*PH*) is collected and compared to a *native* dataset (*P*) that has no heavy atoms bound. Equivalent reflection intensities with the same (*hkl*) indices are subtracted from one another, leaving only the contributions of the heavy atoms. For the subtraction to give useful results, the heavy atom derivative and the native crystal must be *isomorphous*: same space group and their cell dimensions must be very similar, ideally identical, and the structure of the macromolecules in the derivative must not have changed too much by heavy atom binding. Using the *isomorphous differences* as the Fourier coefficients, a *difference Patterson map* is calculated (eq. 24.15). The difference Patterson map will have only a few prominent peaks that correspond to the distances between the heavy atoms, and it is therefore much easier to interpret than the native Patterson map of a whole macromolecule. The positions of the heavy atoms, termed the substructure, are calculated by computers directly from the difference Patterson map. From these positions, the complete structure factors (amplitude and phase) for the heavy atom substructure can be calculated. The phases derived from the substructure allow calculation of starting phases for the macromolecule. In effect, the heavy atoms "replace" the large problem of diffraction by the macromolecule by the smaller problem of finding the substructure. This principle for phasing is also used in variations of isomorphous replacement and in *anomalous diffraction* (Section 24.4.4.2).

BOX 24.3: SOAKING OF MACROMOLECULAR CRYSTALS

The large solvent channels in macromolecular crystals allow molecules of considerable size to diffuse through them. *Soaking* of crystals is an important technique for heavy atom phasing and to determine protein-ligand complexes. To obtain a *heavy atom derivative* of a crystal, the crystals are soaked in heavy atom solutions for a few seconds to many days to allow either covalent or non-covalent binding to the macromolecule. For example, Hg^{2+} and $[PtCl_4]^{2-}$ covalently bind to Cys, His, and nucleobases, whereas I^- (iodide) non-covalently binds to positive charges (Lys, Arg) or hydrophobic sites. An alternative to soaking is to co-crystallize the heavy atom compound with the macromolecule, but soaking has the advantage that many trials with different heavy atom solutions can be carried out in parallel. It is impossible to know whether soaking or co-crystallization have produced a useful derivative without collecting diffraction data. Hence, it is sometimes better to chemically incorporate heavy atoms prior to crystallization (Box 24.4).

Soaking is also used to introduce small molecules such as substrates, products, or inhibitors into crystals. The binding site for such a ligand must not be occluded by crystal contacts, and no large conformational changes must occur upon binding or the order of the crystal might be destroyed. If a protein-ligand complex can be formed *in situ*, its structure can be determined. This is the preferred method in structure-based drug design where hundreds of protein-inhibitor complexes are determined starting from "empty" crystals. Some enzymes that do not undergo large conformational changes during catalysis not only bind their ligand, but also retain their catalytic activity in the crystalline form. By soaking an enzyme crystal with a substrate and stopping the reaction at different time points by cryo-cooling, a series of structures representing the substrate, intermediate, and product complexes may be determined (Section 24.5). Ligand binding and enzymatic activity *in cristallo* are also good indications that the molecular conformation trapped in the crystal is of biological relevance, a question that arises for every crystal structure and normally needs to be answered by independent experiments in solution.

The contribution of the heavy atoms to diffraction and the determination of the phases for the reflections can be visualized by an Argand diagram of the structure factors involved (Figure 24.21). For each reflection (*hkl*) of the native dataset, a structure factor amplitude $|F_P|$ is calculated from the reflection intensity. The phase of this reflection is unknown, so the vector $\boldsymbol{F}_\mathbf{P}$ can have any direction (Figure 24.21). A diffraction experiment with the derivative crystal gives the amplitude $|F_{PH}|$, also with unknown phase for the vector $\boldsymbol{F}_\mathbf{PH}$. The connection between these measurements is the vector $\boldsymbol{F}_\mathbf{H}$, the structure factor of the heavy atom:

$$\boldsymbol{F}_\mathbf{PH} = \boldsymbol{F}_\mathbf{P} + \boldsymbol{F}_\mathbf{H} \qquad \text{eq. 24.16}$$

We know the magnitude and the phase of $\boldsymbol{F}_\mathbf{H}$ from the difference Patterson map. A *Harker construction* geometrically combines the known magnitudes $|F_P|$, $|F_H|$, and $|F_{PH}|$, and the phase for $\boldsymbol{F}_\mathbf{H}$, to estimate the phases of $\boldsymbol{F}_\mathbf{P}$ and $\boldsymbol{F}_\mathbf{PH}$ (Figure 24.21).

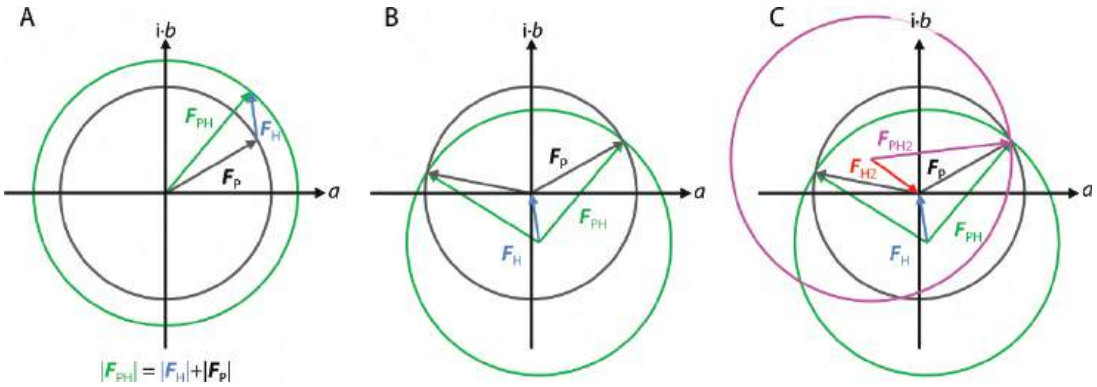


FIGURE 24.21: Harker construction. A: The measured structure factor amplitudes of the protein $|F_P|$ and the derivative $|F_{PH}|$ determine the radii of circles (black and green). The directions, or phases, of $\boldsymbol{F_P}$ and $\boldsymbol{F_{PH}}$ are unknown at first, so these vectors can point to any point on the circumference of their circle. By contrast, both the amplitude and the phase of the structure factor $\boldsymbol{F_H}$ may be calculated from a difference Patterson map (blue arrow), which will now determine the two possible directions of $\boldsymbol{F_P}$ and $\boldsymbol{F_{PH}}$. B: To find these directions, the Harker construction places the vector $\boldsymbol{F_H}$ with its tip at the origin. The circle for $\boldsymbol{F_P}$ (unknown phase; black) has its center at the end of $\boldsymbol{F_H}$, just as in panel (A). The center of the circle for $\boldsymbol{F_{PH}}$ (unknown phase; green) is moved to the beginning of $\boldsymbol{F_H}$. The two circles now intersect at two points, which define the two possible choices of phases for $\boldsymbol{F_P}$ and $\boldsymbol{F_{PH}}$. C: A second derivative (magenta with known $\boldsymbol{F_{H2}}$ in red) resolves the phase ambiguity. This Harker construction is repeated for every reflection (*hkl*).

The Harker construction shows that for each reflection there are two possible phases for $\boldsymbol{F_P}$ (and also for $\boldsymbol{F_{PH}}$). This phase ambiguity is characteristic for single isomorphous replacement (SIR), where we have only a single derivative. The ambiguity can be resolved by repeating the experiment with a second derivative (multiple isomorphous replacement; MIR) where the heavy atom must have bound to a different location on the macromolecule. The structure factor from this derivative (Figure 24.21) can be plotted as a third circle in the Harker construction, and will intersect the other two circles in only a single point, thus resolving the phase ambiguity for $\boldsymbol{F_P}$ (and $\boldsymbol{F_{PH}}$).

24.4.4.2 Anomalous diffraction

Bragg's law relates the angle of incidence θ to the spacing of a set of planes (see Figure 24.15). The same relation holds for an angle of incidence –θ at the same set of planes, but which are now indexed as (*–h–k–l*). Reflection of X-rays from the "underside" of the (*hkl*) planes gives rise to the same reflection intensity, but with opposite phase. This relationship is known as *Friedel's law* (Figure 24.22).

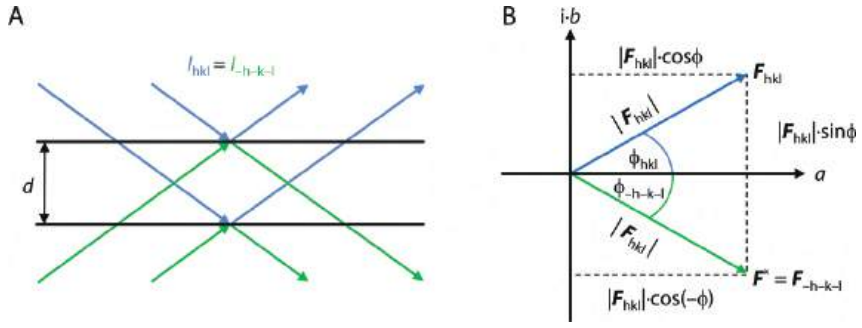


FIGURE 24.22: Friedel's law. A: A set of planes is equally described by the indices (*hkl*) and (*–h–k–l*). Diffraction from this set of planes is possible from angles θ and –θ. Friedel's law states, that the reflection intensities I_{hkl} and I_{-h-k-l} are identical. B: Argand diagram of the structure factors of a Friedel pair, showing that their magnitudes are identical, but the phases have opposite signs.

The reflections I_{hkl} and $I\,I_{-h-k-l}$ are termed *Friedel mates*, and their corresponding structure factors $\boldsymbol{F_{hkl}}$ and $\boldsymbol{F_{-h-k-l}}$ are complex conjugates (denoted by a star; Section 31.11): they have the same amplitude, but opposite phase.

$$\boldsymbol{F}_{\mathbf{hkl}} = \boldsymbol{F}^{*}_{\mathbf{hkl}} \qquad \text{eq. 24.17}$$

This is equivalent to writing:

$$|F_{\mathrm{hkl}}| = |F_{-\mathrm{h}-\mathrm{k}-\mathrm{l}}|$$

$$\phi_{\mathrm{hkl}} = -\phi_{-\mathrm{h}-\mathrm{k}-\mathrm{l}} \qquad \text{eq. 24.18}$$

As a consequence of Friedel's law, the diffraction pattern is *centrosymmetric*, meaning that the distribution of reflections has a center of symmetry just like the Patterson function.

Friedel's law is true for most elements at most X-ray wavelengths. The presence of *anomalous scatterers* in a crystal leads to a breakdown of Friedel's law, so that the intensities of the Friedel mates are not identical any more. Anomalous scattering occurs when the wavelength of the incident X-rays is close to an electronic transition in an atom. In this case, a fraction of the incident photons is absorbed by the atom and re-emitted with a phase shift relative to the incident radiation. This inelastic scattering may reduce the total light intensity by 1–3%, which can be measured and used for phasing. In principle, any atom exhibits anomalous scattering if X-rays of a suitable wavelength are used. The spectral range accessible with synchrotrons (ca. 0.04–0.4 nm) limits the atom types from which useful anomalous signal can be collected to atomic numbers > 16 (sulfur). The anomalous signal becomes increasingly larger for more electron-dense (heavier) atoms.

The presence of anomalous scattering is described by a modified atomic scattering factor:

$$f_{\mathrm{total}} = f_0 + f' + i \cdot f'' \qquad \text{eq. 24.19}$$

f_0 is the wavelength-independent component of the atomic scattering factor (Section 23.2.1; eq. 23.27). f_0 depends on the resolution (and thus on the diffraction angle θ), and it is equal to the atomic number at θ = 0. f' and f'' are the real and imaginary parts of the wavelength-dependent atomic scattering (Figure 24.23). f' is a *dispersive* correction to the normal scattering at a given wavelength. It is usually negative and out of phase by 180° relative to the normal scattering dictated by f_0. The absorption f'' is strongly wavelength-dependent (Figure 24.24) and has a phase shift of 90° relative to f_0, which is signified by the imaginary number i in eq. 24.19. The value of f'' corresponds to the number of electrons contributing to anomalous scattering.

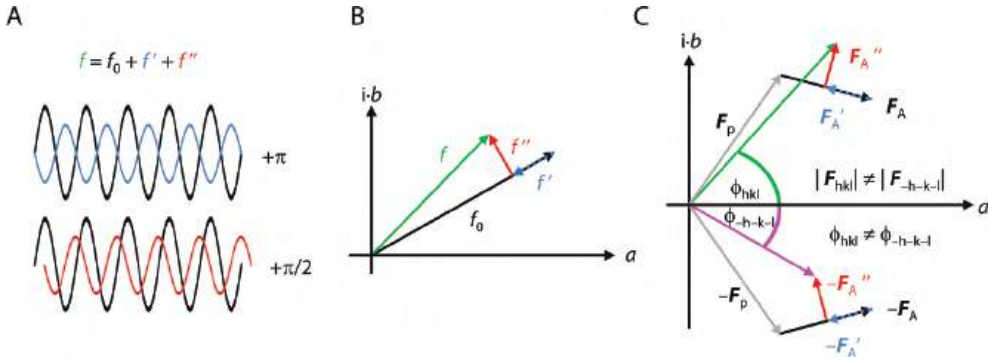


FIGURE 24.23: Modified atomic scattering factor and breakdown of Friedel's law. A: In the presence of anomalous scattering, the atomic scattering factor f_0 is modified by two components that are 90° phase shifted relative to each other. f' is out of phase by 180° relative to f_0, and f'' has a phase advance of 90° relative to f_0. B: Vector representation of the scattering factor components. As f' (blue) is normally negative, the magnitude of f_0 is reduced. Addition of the perpendicular component f'' (red) results in the new scattering factor f (green). C: In the presence of an anomalous scatterer with structure factor $\boldsymbol{F}_{\mathbf{A}}$, the structure factors of the protein $\boldsymbol{F}_{\mathbf{P}}$ and its Friedel mate $-\boldsymbol{F}_{\mathbf{P}}$ (gray) are modified by the same scheme as shown in panel (B). Note that now both the magnitude and the phase of the former Friedel mates (green *versus* magenta) have changed.

To find out whether an anomalous scatterer is present in a crystal, an X-ray fluorescence emission spectrum is recorded, and the absorption and dispersion spectra are calculated (Figure 24.24). The wavelength of maximum anomalous absorption is termed the *absorption edge* or *peak*. The dispersion is maximal at the *inflection point*, which is often only 10^{-4} nm away from the absorption peak. Both f' and f'' are characteristic for each element and the values are tabulated also for many ions and small molecules.

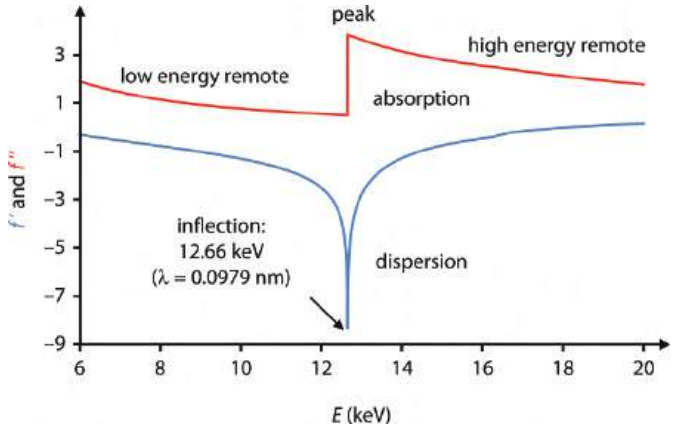


FIGURE 24.24: X-ray absorption. Absorption and dispersion spectra of Se as a function of X-ray energy. The magnitude of f' and f'' depends on the energy (wavelength) of the X-rays. The maximum anomalous signal f'' for Se is 3.8 e^- at $\lambda = 0.0979$ nm (0.979 Å). Atomic scattering factors are tabulated for all elements and many ions.

A diffraction dataset is collected at a wavelength where anomalous diffraction occurs, and the *anomalous differences* of the Friedel mates are used to calculate the positions of the anomalous scatterers from an *anomalous difference Patterson map*. This procedure is similar to the difference Patterson map used to locate heavy atoms in isomorphous replacement, but in anomalous diffraction the differences are calculated using data within a single dataset, not from two different datasets, which eliminates the problem of non-isomorphism. A Harker construction then yields phase estimates for the protein structure factors. In *single wavelength anomalous dispersion* (SAD) the anomalous signal is collected at only one wavelength, often the peak of the elemental absorption with maximum f'' (Figure 24.24), such that the differences between I_{hkl} and I_{-h-k-l} are maximized. SAD has a phase ambiguity similar to SIR, but it can often be resolved by density modification (Section 24.4.5). Recent developments in detector sensitivity and speed of data acquisition enable the weak anomalous signal from the sulfur atoms of Cys and Met for SAD phasing of protein structures to be measured. Sulfur SAD phasing could become a standard technique for *de novo* protein crystal structure determination.

If SAD phasing fails, *multiple wavelength anomalous dispersion* (MAD) can be used. For MAD, diffraction at the *inflection* wavelength with maximum f' is collected in addition to a dataset collected at the peak wavelength. Additional datasets below (*low energy remote*) and above (*high energy remote*) the absorption edge are also often collected and are treated as separate "derivatives" similar to MIR phasing, which usually gives improved phases over SAD, but data collection and all calculations take longer. Care must be taken to avoid radiation damage, which easily changes the reflection intensities far more than the magnitude of the anomalous signal. Anomalous phase information can be readily combined with phases from single and multiple isomorphous replacement, termed *SIRAS* and *MIRAS*. The result of a phasing experiment that includes anomalous signal can be an electron density map of stunning quality (see Figure 24.27).

BOX 24.4: COVALENT INTRODUCTION OF HEAVY ATOMS AND ANOMALOUS SCATTERERS INTO MACROMOLECULES.

Many heavy atoms that are soaked into crystals for MIR phasing also have anomalous signal at the wavelengths to which synchrotrons can be tuned. However, often the soaked heavy atoms fail to bind to the protein, which is only found out after lengthy data collection and evaluation. An alternative is to covalently bind the anomalous scatterer to the macromolecule (Figure 24.25). 2'-bromo and 2'-seleno nucleotides can be included in DNA and RNA during synthesis, provided there is enough space at the 2'-deoxy position (DNA) or the 2'-OH group is not involved in tertiary interactions (RNA). U and T have also been replaced by 5-bromouracil.

The most popular method for proteins is *Se-Met modification*. An *E. coli* culture in minimal medium without Met but including Se-Met will incorporate Se-Met into proteins. Alternatively, the culture can be grown in rich medium, and Met biosynthesis may be inhibited prior to induction of the culture by addition of Ile/Leu/Phe/Lys/Thr at high concentrations. The remaining Met is quickly depleted by protein biosynthesis, followed by incorporation of the added Se-Met. This approach is also used for introducing isotopes into proteins (see Box 22.6). The content of the incorporated Se-Met may be analyzed by mass spectrometry. It is useful to know the percentage of Se-incorporation before the time-consuming process of crystallization and data collection. A slight disadvantage of Se-Met is its hydrophobicity and slightly larger volume compared to Met. Se-Met proteins therefore tend to be less soluble and are sometimes less stable than their natural Met counterparts.

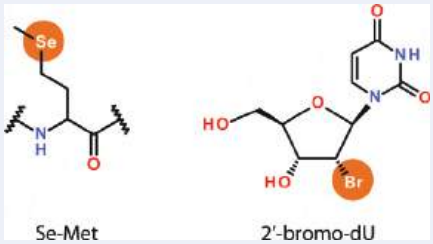


FIGURE 24.25: Seleno-methionine (Se-Met) and 2′-bromo-uridine (2′-bromo-dU).

24.4.4.3 Molecular replacement

From a conceptional point of view, molecular replacement is the simplest way of obtaining phases. It is also the most widely used method for phasing: over 70% of all structures in the PDB have been determined using molecular replacement. The phases are calculated from a similar structure, the *search model*, after placement of that model in the unit cell of the observed diffraction data. "Placement" means that the search model needs to be rotated and translated to become superimposed on the unknown structure. Computationally, this is a six-dimensional problem because a superposition of two molecules is described by three rotation angles about the axes of the coordinate system, and three translations along the axes. The position A' of the atoms in the target molecule is given by their position A multiplied by a rotation matrix $\boldsymbol{R}$ plus a translation vector $\boldsymbol{T}$:

$$A' = A \cdot \boldsymbol{R} + \boldsymbol{T} \qquad \text{eq. 24.20}$$

The problem of molecular replacement is solved when $\boldsymbol{R}$ and $\boldsymbol{T}$ have been found. Using the properties of the Patterson function (Section 24.4.3), the molecular replacement calculation can be split into separate rotation and translation steps (Figure 24.26). Two three-dimensional searches are computationally much more efficient than a single six-dimensional search.

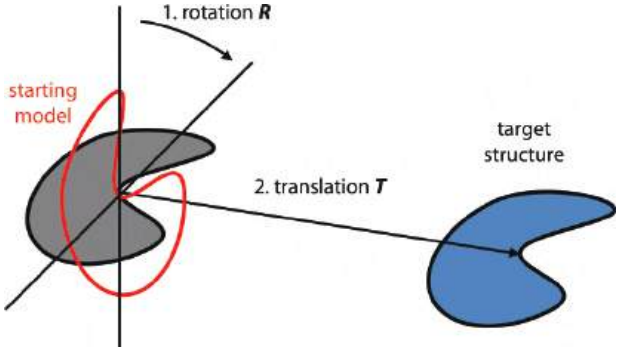


FIGURE 24.26: Molecular replacement. The starting model is first rotated by Patterson function comparison (rotation function). The translation function places the rotated model at the correct location in the target unit cell.

The rotation search makes use of the Patterson functions of the search model, P_{calc}, and the diffraction data, P_{obs}. We recall that the Patterson function can be calculated in the absence of phases and that it represents a set of distance vectors $\boldsymbol{u}$ between atoms in a molecule. If we rotate one Patterson function in three dimensions with respect to the other, we should be able to find a relative orientation where the sets of vectors match closely. This calculation is done by a *cross-rotation function* RF_R, which multiplies the two different Patterson functions after one of them (P_{calc}) has been rotated by an angular increment with respect to the other (P_{obs}).

$$RF_{\mathbf{R}} = \int P_{obs}(\boldsymbol{u}) \cdot P_{calc}(\boldsymbol{R} \cdot \boldsymbol{u}) \cdot \mathrm{d}u \qquad \text{eq. 24.21}$$

The angular increment is encoded in the matrix $\boldsymbol{R}$. If the agreement between the Patterson functions after rotation is high, RF_R has a maximum and the rotation of the search model with respect to the target structure is found. The maximum vector magnitude u for the integration should correspond to the diameter of the search model in order to limit the calculation to predominantly intramolecular distances. This value can be estimated from the sequence or molecular mass of the target molecule, assuming that it is roughly spherical. If many intermolecular vectors enter the calculation by choosing a too large value for u, the cross-rotation function may get too noisy to find the correct rotation solution. This is a problem when searching for very elongated molecules made of α-helices such as leucine zippers, helix bundles, and filaments.

The next step is to find the position of the correctly oriented model in the unit cell by use of a *translation function* TF_t. Like the rotation function RF_R, TF_t is the product of the two Patterson functions P_{obs} and P_{calc}. Now, P_{calc} is the Patterson function calculated from the search model after the rotation matrix found in the rotation step has been applied. P_{calc} is then translated on a three-dimensional grid throughout the entire unit cell by a vector $\boldsymbol{T}$.

$$\mathrm{TF}_{\mathbf{t}} = \int P_{obs}(\boldsymbol{u}) \cdot P_{calc}(\boldsymbol{u} + \boldsymbol{T}) \cdot \mathrm{d}u \qquad \text{eq. 24.22}$$

Upon translation of P_{calc}, the intramolecular vectors stay the same while the intermolecular vectors change. Modern implementations of the translation function subtract the set of intramolecular vectors to remove noise from the calculation and use only the intermolecular vectors to decide on the correct position of the search model. A translation vector $\boldsymbol{T}$ that leads to superimposition of P_{obs} and P_{calc} produces a maximum in TF_t. Many solutions are usually found, and the next step is to distinguish the correct solution from the best wrong solution. A *packing analysis* helps here: the symmetry operators of the space group are applied to the rotated and translated search model to fill the unit cell, and the number of overlaps between symmetry-related molecules is counted. The best solution should produce a minimum number of overlaps, maybe between loop regions that differ between the search model and the correct structure. This is checked by calculating an electron density map

using the phases from the molecular replacement solution and the amplitudes $|F_{obs}|$ from the measured reflections. If the electron density contains information that was not present in the search model, such as different loop conformations or a bound ligand, the molecular replacement solution is very likely correct. False-positive molecular replacement solutions may arise due to *model bias*, where a search model refined to high resolution is compared to high-resolution diffraction data. Molecular replacement programs try to reduce the bias by either cutting the resolution to 0.4–0.5 nm or using additional quality criteria to the height of the rotation and translation functions, including cross-correlation coefficients (Section 31.12) and maximum likelihood targets (Section 24.4.6).

The search models used for molecular replacement might differ by just a few atoms from the true structure, for instance in variants of a few side chains or when a different ligand is bound. Such structure determinations are usually straightforward. If the crystals from search model and target are isomorphous, molecular replacement is not even needed but a simple rigid-body refinement (Section 24.4.6) of the search model against the diffraction data is sufficient. This approach is sometimes called *molecular substitution* and finds application in high-throughput structure determination such as ligand screening and drug design.

Sometimes, however, the closest known search model is so different from the true structure that molecular replacement fails. The minimum structural similarity between a search model and the true structure is preferably less than 0.15 nm rmsd (see eq. 31.2), but this is not known *a priori*. The sequence identity is therefore used as a guide to judge the likelihood of success for a given search model. More than 40% identity should give a clear solution, unless there are large conformational differences between search model and the unknown structure. In this case sub-domains of the search model can be used one after the other. Dividing the search models into sub-domains is frequently needed for antibody Fab structure determination because the four Ig-domains constituting the Fab (Figure 24.5) can move with respect to each other.

The minimum sequence identity between the search model and the target structure is about 20%, but low resolution of the data and a large number of molecules in the asymmetric unit that have to be searched for may require better search models. Search models can be improved *in silico* by removing non-conserved parts identified from multiple sequence alignments (see Section 18.2.2). If search models of < 30% sequence identity are successful in determining the structure, it is because the three-dimensional fold is more conserved than expected. The conservation of three-dimensional folds despite large differences in sequence is the basis of novel developments in molecular replacement, which promise to determine many new crystal structures without the need for isomorphous and anomalous phasing (Box 24.5).

BOX 24.5: MOLECULAR REPLACEMENT AS THE ULTIMATE PHASING TOOL?

Although the number of crystal structures in the PDB keeps growing at an almost exponential rate, fewer and fewer novel folds are added to the database (Section 18.3.4). It is likely that the number of different folds in nature is limited, and that each fold is already represented by a crystal structure in the PDB. This raises the possibility that any new structure could be phased by molecular replacement using search models cleverly pieced together from known structures. Two promising approaches have been developed, one that uses predicted models and another that uses secondary structure and super-secondary structure elements as search models. Secondary structures of nucleic acids are often successful in determining RNA and DNA structures, where DNA is usually less problematic due to its more conserved structure compared to RNA. In the long run, the techniques promise to determine any crystal structure by molecular replacement using only the information we already have from previous structures or sequence alignments.

We have seen in Section 20.3 how large sequence alignments can lead to quite accurate models of proteins. If the coordinate uncertainties of the models are < 0.1 nm compared

(*Continued*)

to the (unknown) true structure, the models may be used directly as search models for molecular replacement. If, however, the target structure has large conformational differences compared to the model, a different approach is needed.

ARCIMBOLDO and BORGES (chango.ibmb.csic.es/ARCIMBOLDO) are two programs that do not use homology models for the search but thousands of secondary and super-secondary structure elements derived from the PDB. The required accuracy for a successful search model is below 0.06 nm rmsd, explaining the huge number of searches that have to be conducted. The approach currently works best with α-helices as these show the least structural variation and have high electron density due to the helical arrangement of the residues. β-sheets and loops are more extended and have more variable structures, making their identification more difficult. A few dozen structures, predominantly α-helical, have been determined so far. The diffraction data should extend to at least 0.25 nm resolution, and the asymmetric unit should have less than 400 residues.

24.4.5 Electron density and model building

After approximate phases have been obtained by any of the methods described in the previous Sections, a first electron density map is calculated (see eq. 24.10) using the phases and the measured structure factor amplitudes $|F_{obs}|$. The next step on the way to the crystal structure is interpretation of the electron density in terms of atoms by threading a polypeptide or oligonucleotide chain through the density. However, in many cases the quality of the initial electron density map is too poor for interpretation. *Density modification* procedures improve the phases, and thus the electron density maps, by including additional information, as described in the following.

Solvent flattening relies on the fact that the solvent between the macromolecules has a different (usually lower) electron density than the macromolecules (see Table 23.3). An envelope around a volume corresponding to the macromolecule is identified, and the electron density outside the envelope is set to zero or to a fixed value (*e.g.* 330 e^-/nm^3 for water). From these modified maps, new phases are calculated by inverse Fourier transformation. With the improved phases and the experimental structure factor amplitudes $|F_{obs}|$, a new electron density map is calculated that is often easier to interpret.

Histogram matching is based on a feature used in image processing and photography that assumes a rather constant or only gradually changing value across an object (color in images, electron density in molecules). The electron density should be rather constant within molecules, and if the initial electron density map has large discontinuities, these may be smoothed. For this purpose, the electron density map is binned into histograms and compared to the expected *electron density histogram* of proteins and nucleic acids. The reference histograms of proteins and nucleic acids are remarkably similar and rather independent of sequence or structure. Adjusting the density distribution of an initial map to match the expected values yields new and improved phases.

Non-crystallographic symmetry (NCS) averaging can be applied when at least two independent copies of the same molecule are present in the asymmetric unit. These molecules should have equal electron densities at corresponding sequence positions. The method requires an NCS operator for each pair of molecules for which the electron density should be averaged. NCS averaging can spectacularly improve the quality of the electron density maps, particularly when many NCS copies are present such as in highly symmetric virus structures. In essence, the more NCS copies, the more restrained are the phases due to averaging.

If the resulting electron density map is of sufficiently high quality, the outline of the macromolecules are visible (also termed a *solvent boundary*), and secondary structure elements can be recognized (Figure 24.27). In favorable cases, the main chain can be traced and many side chains can be assigned just on the basis of electron density (Figure 24.27). Usually, however, knowledge of

the sequence is required to assign all side chains. In general, the resolution of the electron density map is determined by the measured diffraction data and determines the level of detail (Figure 24.27): α-helices are the most compact secondary structure elements and are visible as rods in electron density maps down to 0.5 nm resolution. β-sheets become discernible at around 0.4 nm resolution. Tracing of the main chain and some flexible loops requires around 0.3 nm resolution. Side chains, especially the large aromatic Trp, Phe, Tyr, and His, also become well defined at this resolution. At 0.2 nm, most side chains and *cis*-peptides are clearly visible. Finally, atomic resolution structures ($\leq$ 0.1 nm) allow visualization of individual atoms (Figure 24.27). Since hydrogen atoms contain only a single electron, they are normally not visible in electron density maps. Neutron crystallography (Section 24.6) is the preferred method to study hydrogens by diffraction methods.

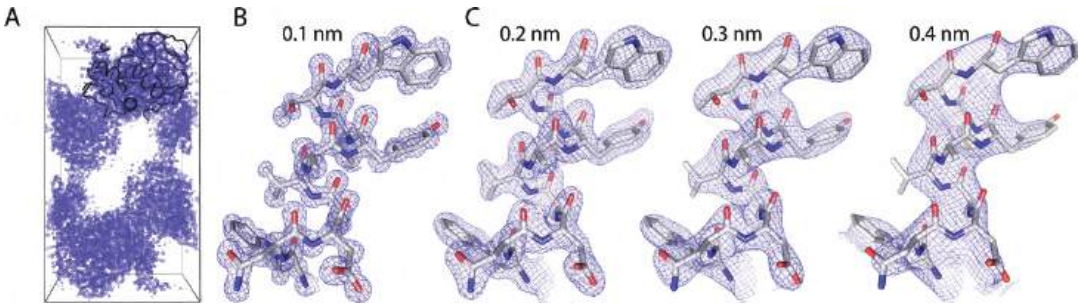


FIGURE 24.27: Electron density after density modification and effect of resolution. A: The clear solvent boundary indicates accurate phases. One molecule of the unit cell is outlined by a black line. B: Magnification of a part of the density in (A) showing an α-helix with well resolved side chains at 0.1 nm resolution. C: The electron density map in (B) was recalculated to lower resolutions of 0.2 nm, 0.3 nm, and 0.4 nm. Note the decreasing quality of side-chain density. The main-chain trace of the α-helix is still visible at 0.3 nm resolution. At 0.4 nm resolution, the α-helix resembles a rod. The information content of electron density strongly depends on the quality of the phases. In the examples shown, the phase error is small (~30°).

A particularly useful kind of map for model building is the *difference map*. It is calculated with Fourier coefficients $|F_{obs}|-|F_{calc}|$, where $|F_{obs}|$ comes from the measured data and $|F_{calc}|$ and ϕ_{hkl} are calculated from the model:

$$\Delta\rho_{xyz} = \frac{1}{V} \cdot \sum_{hkl} \left| |F_{obs}| - |F_{calc}| \right| \cdot e^{i\phi_{hkl}} \cdot e^{-2\pi i \cdot (hx+ky+lz)} \qquad \text{eq. 24.23}$$

Since difference maps require $|F_{calc}|$ and ϕ_{hkl} from a model, they are usually calculated after model refinement (Section 24.4.6). Difference maps are positive in regions where atoms are missing, and negative in regions where too many electrons (atoms) are present in the model. Hence, a difference map provides important hints where and in which direction the model needs to be corrected to better describe the electron density (Figure 24.28). A second difference map using $2|F_{obs}|-|F_{calc}|$ as the Fourier coefficients is also calculated. This map weights F_{obs} twice as strongly as F_{calc} to emphasize the measured diffraction data (Figure 24.28) and represents the overall electron density of the molecule.

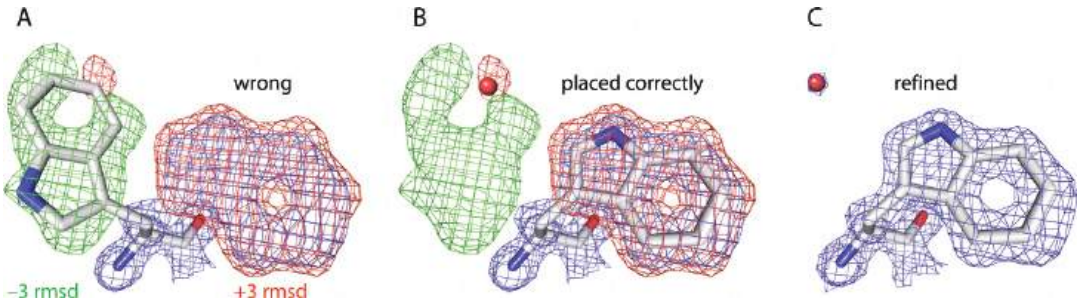


FIGURE 24.28: Difference electron density. A: The Trp side chain is in the wrong orientation as shown by the negative difference density (green). The correct placement would be in the positive difference density (red), which also shows a water molecule (red sphere) that occupies some of the space taken by the wrong Trp side chain. B: Corrected model before refinement. Another rotamer of Trp was chosen, and a water molecule was assigned to the isolated difference density on the top left (its binding partners are not shown). C: Maps after refinement. There is no difference density any more, just $2|F_{obs}|-|F_{calc}|$ density (blue).

The electron density map of a unit cell represents the average electron density of all unit cells that have contributed to the diffraction data. Thus, the electron density contains information on a huge number of molecules, which may differ slightly in their conformations. Sometimes, and especially at high resolutions of < 0.2 nm, the electron density maps show more than one conformation for part of a molecule. The most frequently encountered case is side chains in alternate conformations, which can be modeled by assigning different *occupancies* to the conformations (Figure 24.29). The occupancy is the fraction of each conformation throughout the crystal. For covalently bonded atoms, the sum of all occupancies must be unity, while non-covalently bound molecules, such as water, metal ions, cofactors, and ligands, may have occupancies < 1 because they are not necessarily present in all unit cells.

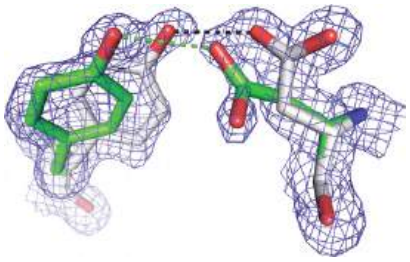


FIGURE 24.29: Alternate conformations. These Tyr and Asp side chains have alternate conformations that depend on each other. The conformations in gray color form a hydrogen bond (dashed black line), as does the pair shown in green. While the main-chain atoms are almost unchanged and have occupancies of unity, the side-chain atoms were duplicated and assigned occupancies of 0.7 (gray) and 0.3 (green).

The first electron density map does not usually allow building of a complete and error-free model. The initial model is refined against the diffraction data (Section 24.4.6) to obtain improved phases that are used to calculate new difference maps, which are then used in the next round of model building. This iterative cycling between model building and refinement is performed until the resulting electron density maps do not show novel features that should be included in the model.

24.4.6 Model refinement and validation

Refinement modifies the coordinates and positional uncertainties (*B*-values; see later in this Section) of the atoms in the model to match the diffraction data as closely as possible, while at the same time covalent and non-covalent interactions are maintained with plausible geometry (Sections 17.2.1 and 17.2.2). These two aspects of refinement can be expressed by an energy function that is a weighted (w_{xray}) sum of geometric ($E_{geometry}$) and X-ray diffraction (E_{xray}) energy terms:

$$E_{\text{tot}} = E_{\text{geometry}} + w_{\text{xray}} \cdot E_{\text{xray}} \qquad \text{eq. 24.24}$$

During the refinement E_{tot} is minimized. The geometric energy includes the *restraints* on the structure: energy terms for bonds, angles, torsions, planes, chiralities, as well as ionic, dipolar, and van der Waals interactions (see eq. 20.1). The energy function can also include any other prior information that is independent of the diffraction data, such as non-crystallographic symmetry, if present. The purpose of including restraints is to decrease the number of parameters that are fitted during refinement, *i.e.* to increase the data-to-parameter ratio. For example, in the absence of geometric restraints, a typical crystal with 43% solvent content that diffracts to a resolution of 0.25 nm corresponds to a data-to-parameter ratio of unity, which is too small for finding the global minimum of the energy function. For high-resolution diffraction data with many reflections, the restraints are less important and the X-ray energy term E_{xray} may be weighted higher by increasing w_{xray}.

Traditionally, E_{xray} was of the form

$$E_{\text{xray}} = \sum_{\text{hkl}} \left(|F_{\text{obs}}| - k \cdot |F_{\text{calc}}| \right)^2 \qquad \text{eq. 24.25}$$

where $|F_{\text{obs}}|$ and $|F_{\text{calc}}|$ are the structure factor amplitudes of the reflection (*hkl*), and k is a relative scale factor between F_{obs} and F_{calc}. This form of E_{xray} is the *least squares target*, which in older refinement algorithms was minimized using the conjugate gradient minimization procedure (Section 20.1.2). The currently most successful formulation for the X-ray energy term is the *maximum likelihood* (ML) target,

$$E_{\text{xray}} = \sum_{\text{hkl}} \frac{1}{\sigma_{\text{ML}}^2} \left(|F_{\text{obs}}| - \langle |F_{\text{obs}}| \rangle \right)^2 \qquad \text{eq. 24.26}$$

where $\langle |F_{\text{obs}}| \rangle$ is the expected value for $|F_{\text{obs}}|$, and $\sigma_{\text{ML}}{}^2$ is the variance. Maximum-likelihood methods determine the probability of making a certain set of measurements (the diffraction dataset) given a molecular model, the diffraction data, and error estimates for the model and the data. The main problem is to estimate the unknown errors of the (incomplete) model. It can be shown that the probability distribution of a structure factor from the model, F_{calc}, depends on the parameter σ_{A}, which can be estimated from the experimental data. σ_{A} essentially represents the fraction of the structure factor that is expected to be correct, and its value ranges from zero (completely incorrect) to unity. Together with the known $|F_{\text{obs}}|$ and the $|F_{\text{calc}}|$ from the model, σ_{A} allows one to calculate the expected values for $\langle |F_{\text{obs}}| \rangle$ and the corresponding variance $\sigma_{\text{ML}}{}^2$ in eq. 24.26. The ML target has proven so successful that all modern crystallographic refinement programs have it implemented.

As a result of the energy minimization during refinement, the difference between F_{obs} and F_{calc} is minimized. The crystallographic *R*-factor R_{cryst} is a measure of the residual magnitude of this difference:

$$R_{\text{cryst}} = \frac{\sum_{\text{hkl}} \left| |F_{\text{obs}}| - |F_{\text{calc}}| \right|}{\sum_{\text{hkl}} |F_{\text{obs}}|} \qquad \text{eq. 24.27}$$

The sum in the denominator of eq. 24.27 normalizes the sum of differences in the numerator, so R_{cryst} adopts values < 1. A perfect match between model (F_{calc}) and data (F_{obs}) would have $R_{\text{cryst}} = 0$. At the outset of refinement, R_{cryst} values of 0.5 (50%) are not unusual, but will eventually drop to values of $< 25\%$, maybe $< 15\%$ for a well-refined structure. The final value of R_{cryst} depends on the resolution and quality of the data, and the stereochemical quality of the model built, but *R*-factors for macromolecular structures almost never approach zero because of measurement errors in the diffraction data and sub-optimal energy functions.

Sometimes, R_{cryst} is artificially reduced by refinement programs without any actual improvement of the molecular model. This corresponds to a decrease in the energy function by accumulation of

systematic errors in the model, known as *overfitting*. The *free R-factor* R_{free} avoids this problem: R_{free} is calculated in the same manner as R_{cryst} (eq. 24.27), but from a test set of ideally > 1000 reflections that are never used in refinement. This test set is "free" from bias of the model for the diffraction data, which is introduced by any refinement procedure. R_{free} therefore represents a much more reliable quantity to judge the overall progress of refinement than R_{cryst}, but at the cost of not using all available data. As a rule of thumb, R_{free} should roughly correspond to the resolution of the diffraction data in nm, *e.g.* a structure determined to 0.25 nm resolution should have $R_{free} \leq 25\%$. Since a certain amount of overfitting is introduced by any refinement procedure, R_{cryst} will always be lower than R_{free} but should not differ by more than a few (3–8) percentage points.

At the beginning of the refinement the model is likely very incomplete and contains many errors, requiring methods such as rigid-body refinement and simulated annealing (Section 20.1.3) that apply large changes to the coordinates. While simulated annealing usually moves secondary structure elements, loops, and side chains, rigid-body refinement is a six-dimensional search (3x rotation, 3x translation) for the best fit of whole domains to the diffraction data. Later refinements with a more correct model require smaller adjustments of atomic coordinates and *B-values*. As the electron density is a time and space average, the crystal structure also represents a time and space average. The *B*-value is used to approximate the observed "blurring" of the electron density due to the combined effect of atomic thermal vibrations (dynamic disorder) and small differences in the atomic positions between crystallographically "equivalent" molecules (static disorder). Both dynamic and static disorder are expressed together as the isotropic (spherical movement assumed) uncertainty u of the center of gravity of an atom in any particular direction x:

$$B_{iso} = 8\pi^2 \left\langle u_x^2 \right\rangle \qquad \text{eq. 24.28}$$

The radial displacement u_r, *i.e.* the total distance from the mean position, is calculated from the mean square radial displacement $\langle u_r^2 \rangle$, which is three times the mean square displacement $\langle u_x^2 \rangle$. Hence,

$$B_{iso} = \frac{8}{3}\pi^2 \left\langle u_r^2 \right\rangle \qquad \text{eq. 24.29}$$

Regardless of which view is taken, displacement in a particular direction or radial, the larger the *B*-value, the less well defined is the atom position. When the unit of u is given in Å, the unit of the *B*-value is Å^2. The atomic resolution crystal structures of small molecules usually have tiny *B*-values around 1–2 Å^2, which reflect purely dynamic disorder, *i.e.* represent the uncertainty of the atom position as a harmonic oscillation. Macromolecules in crystals display much more static and dynamic disorder because they are packed less tightly than small molecules. The corresponding crystal structures have mean *B*-values of 20–80 Å^2. According to eq. 24.28, a *B*-value of 79 Å^2, which is not unusual for macromolecular structures, corresponds to an uncertainty in atom position of 1 Å (100 pm) in each direction. Because this uncertainty is on the order of the C-C bond length (154 pm), it is clear that such *B*-values must describe to a large extent static disorder.

Instrument parameters such as X-ray beam divergence, detector noise fluctuations, and vibrations of the crystal during data collection can also influence the *B*-value. Therefore, it is difficult to compare *B*-values across different crystal structures, but within a single crystal structure, the *B*-value is useful to compare different parts of the same molecule: a stretch of elevated *B*-values relative to the average could indicate high mobility of this region (Figure 24.30). Crystallographic *B*-values are not the best method to assess flexibility, but if several structures of the same molecule in different packing environments show the same trend, it may be assumed with some confidence that these regions are also flexible in solution.

If the resolution of the diffraction data is ≤ 0.15 nm, the isotropic *B*-value B_{iso} can be replaced by an *anisotropic B-value* B_{aniso}. Instead of describing the atomic uncertainty by a single parameter as identical in all directions, B_{aniso} uses a set of six parameters for the orientation and extent of an

ellipsoid in which the atom may be found (Figure 24.30). Anisotropic refinement is also possible for lower resolution structures by assigning anisotropic properties to whole domains or secondary structure elements instead of individual atoms. This refinement technique is known as TLS (Translation, Libration, Screw) refinement.

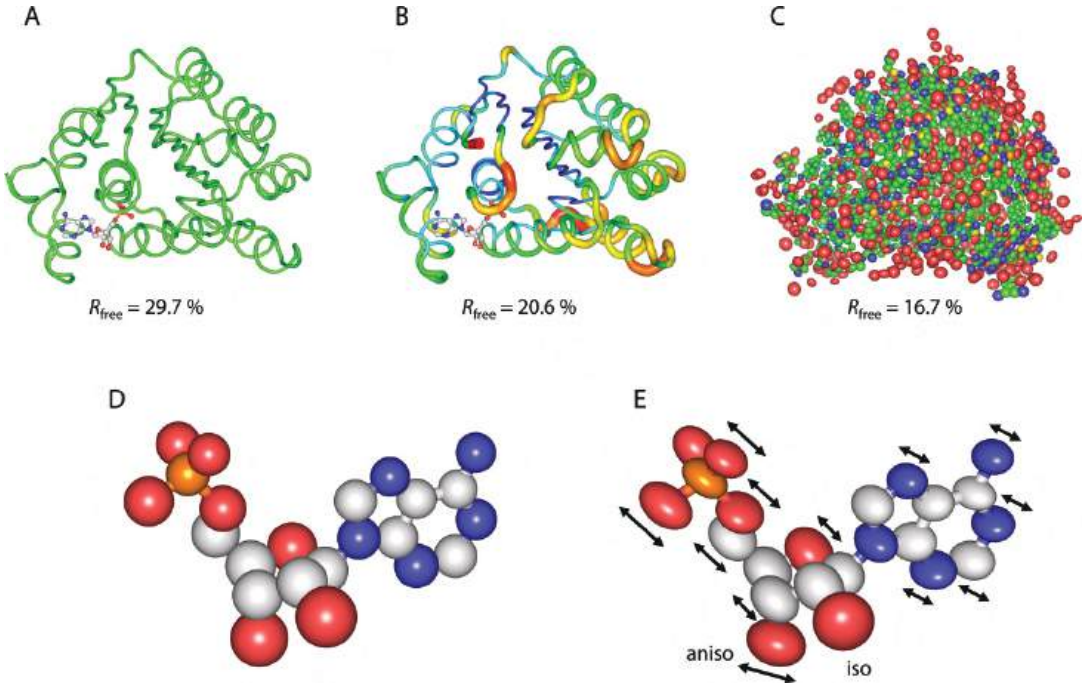


FIGURE 24.30: *B*-value refinement. The effect of isotropic and anisotropic *B*-value refinement is shown for a 0.12 nm resolution structure with the ligand AMP bound (PDB-ID 2gxq). A: No *B*-values are refined, every atom has the same *B*-value of 15.2 Å^2, and R_{free} is almost 30%. B: Isotropic *B*-value refinement reduces R_{free} to 21% and shows surface areas of potential flexibility (red and thick tubes). Coloring is from blue (low *B*-values) to red (highest *B*-values). C: Anisotropic *B*-value refinement reduces R_{free} to about 17%. The atoms are shown as thermal ellipsoids. D: Isotropic *B*-values for the AMP ligand from (B) show higher apparent flexibility of the ribose part (larger spheres) than the adenine nucleobase. E: Anisotropic *B*-values for AMP from (C) reveals that the phosphate and adenine groups seem to oscillate as individual groups. The adenine atoms oscillate perpendicular to the plane of the nucleobase. Some atoms stay isotropic, for example the 2′-hydroxyl group (labeled *iso*), indicating that it has no preferred oscillation direction.

When after a few cycles of building and refinement certain quality criteria such as reasonable *R*-factors and geometry are met, the iterative process may be aborted and the model considered final and ready for biological analysis. There are basically two classes of quality checks, those based purely on coordinates (does the model make sense?), and those that refer to diffraction data (how well does the model correspond with the diffraction data?). The parameters expressing the model quality can be global or local. Among the most important ways to test the quality of a crystal structure are R_{free}, the Ramachandran plot, and electron density.

R_{free} is one of the most frequently used global quality indicators that refer to diffraction data. Other global quality indicators, such as the rmsd values for bonds, 1,3-angles, and torsion angles are not good descriptors for structure quality since they are used as targets in refinement: even a wrong crystal structure can have quite acceptable geometry. Provided that the main-chain torsion angles have not been restrained to the Ramachandran values during refinement, the Ramachandran plot (see Figure 18.21) is an invaluable tool to pinpoint residues in high-energy conformations, often located at enzyme active sites, metal binding sites, or as *cis*-peptides. If the electron density agrees with the peculiar geometry, it is a genuine feature, not an error.

The best local quality indicator for a crystal structure is visual inspection of the electron density map. Deposition of diffraction data in the PDB is mandatory for new structures since 2008,

so maps can be calculated for "newer" crystal structures and displayed in programs such as Coot (github.com/pemsley/coot), CCP4mg (ccp4.ac.uk), VMD (www.ks.uiuc.edu/Research/vmd), Pymol (pymol.org), or ChimeraX (www.cgl.ucsf.edu/chimerax). Some programs, Coot for example, allow input of a PDB-ID and directly contact the PDB or the automated structure re-refinement database PDB_REDO (www.cmbi.ru.nl/pdb_redo), so electron density can be judged with minimal effort. The advantage of inspecting the electron density is that most researchers are interested in local aspects of a crystal structure (an active site, a ligand pose, an interaction surface), and this is directly represented by electron density. If, for example, the electron density for a ligand is weak compared to the density of the surrounding protein atoms, the ligand may have high mobility when bound to the protein or an occupancy less than unity or the whole ligand may simply be a modeling error. Similar reasoning applies to metal ions, which should have the expected geometry and distances to their coordinating atoms (Table 18.3). It is entirely possible to obtain an atomic resolution (< 0.1 nm) diffraction dataset but observe uninterpretable electron density for a certain set of atoms. A numeric quantity for local electron density quality is the real-space correlation coefficient, which can be calculated on a per-atom basis and should be close to unity.

24.5 TIME-RESOLVED CRYSTALLOGRAPHY

For obtaining kinetic information from crystals, data collection needs to be faster than the reaction followed in the crystal. Fast data collection is also essential to minimize radiation damage, a strategy used in serial crystallography at synchrotrons and XFELs.

There are two approaches to obtain time-resolved high-resolution structural information using crystallography: fast data collection of an ongoing reaction (continuous) and stopping a synchronized reaction *in cristallo* at different time points followed by offline analysis (discontinuous). In the discontinuous approach, a slow chemical reaction is started throughout an enzyme crystal by soaking a substrate into a number of crystals. The reaction is stopped at different time points by cryo-cooling a single crystal, and its structure is determined. The crystal structure will report on the progress of the reaction by way of accumulation of intermediates and products. An alternative way of synchronizing the chemical reaction in crystals is laser flash photolysis of suitable systems (Section 28.3). Reversible *cis*/*trans* isomerization of the double bond in the chromophore of photoactive yellow protein and cleavage of the Fe-CO bond in the heme group of carbonmonoxy-myoglobin are reactions inducible within the cryo-cooled crystal by a UV-laser pulse. Laser flash photolysis can also trigger chemical reactions by activation of an inactive precursor, a *caged compound* (see Figure 28.6; Figure 28.7) including enzyme substrates or masked metal ions. Until fast pixel detectors were developed, these continuous reactions triggered by laser pulses were usually followed by *Laue diffraction*. In contrast to conventional monochromatic crystallography, Laue diffraction forgoes the monochromator at the synchrotron beamline and just uses a mirror for a high-energy radiation cut-off of the undulator harmonics. This *pink beam* has a relative photon energy spread $\Delta E/E \approx 0.05$ and yields ca. 100 times larger flux than monochromatic radiation. The resulting diffraction images contain many more reflections than a single-wavelength exposure because the Bragg condition is fulfilled for several wavelengths at the same time, leading to superposition of several diffraction patterns on a single image. Hence a dataset can be collected much more quickly compared to a single-wavelength experiment but the data is also usually less complete: reflections from different wavelength diffraction patterns that overlap on the same detector pixels need to be excluded from the final dataset. The time resolution for early Laue diffraction studies was in the millisecond to nanosecond range. Over the years the time resolution approached the length of a single bunch in the storage ring, or about 100 ps. Even higher time resolutions in the femtosecond range are possible with XFELs (see Section 24.6). Also, in order to avoid the complications of the Laue method, serial crystallography at room temperature is becoming increasingly popular. Kinetic information from serial crystallography is gained by synchronization of a reaction in a slurry of micro-crystals and injecting droplets at constant velocity and at different time points from reaction initiation perpendicularly across the X-ray beam.

24.6 SERIAL CRYSTALLOGRAPHY AND X-RAY FREE ELECTRON LASERS

The strong radiation produced by third-generation synchrotrons and XFELs allows analysis of sub-µm crystals. Micro-crystals are often easier to produce than larger crystals, especially for membrane proteins. Sometimes they even grow *in vivo* during protein production, confirming the general principle that the crystalline state is a thermodynamically stable state. However, their tiny volume renders them very susceptible to radiation damage (see Box 24.2), especially if room temperature measurements are required. The advantage of room temperature diffraction data is that the B-values of the atoms in the resulting crystal structure may contain more biologically relevant information on protein flexibility. Serial crystallography avoids radiation damage by "measurement before destruction": the data are collected so quickly (in femtoseconds) that any radiation damage setting in (after ca. 100 ms) does not affect the diffraction pattern. In *serial femtosecond crystallography*, a few thousand small crystals are presented one after the other to the X-ray beam, with the hope of collecting a single diffraction image of each crystal. A complete diffraction dataset consists of thousands of images, each from a different crystal. Serial crystallography was first developed to allow data collection on XFELs, but later the principles were adapted to strong synchrotron beamlines.

There are two main methods to deliver crystals to the X-ray beam in serial crystallography, by droplets and, less frequently, by conveyor belt. In the droplet approach, the crystals are suspended in a viscous buffer and injected into the beam from a capillary or syringe. Aqueous buffers provide higher flow rates while lipidic cubic phases (see Box 24.1) are suitable for higher viscosities, lower flow rates, and the study of membrane proteins (Figure 24.6). The aim is to dilute the suspension of crystals and adjust the viscosity such that each droplet holds a single crystal and has a high chance of being hit by X-rays. The conveyor belt method places small droplets containing crystals on a tape that moves perpendicular to the XFEL beam. Tapes have the advantage that the drop rests and moves just in one dimension, increasing the hit rate compared to the injection method. In practice it is not uncommon for either method to find droplets empty or to contain several crystals, resulting in blank images or overlapping diffraction patterns. Timing in both delivery methods is crucial. If the X-ray beam does not hit a crystal, it is lost nonetheless. Typical hit rates are small, which explains the high sample consumption by such experiments of often several gram of crystals. Fixed-target serial crystallography, where the crystals are placed in small depressions of silicon nitride "chips", promises increased hit rates with concomitant smaller sample consumption.

If the drop delivery frequency coincides with the pulsed synchrotron or XFEL light, the droplet is disintegrated by the beam, but a diffraction pattern is collected before that by ultrafast direct photon detectors (Section 24.4.3). Crystals suspended in droplets are free to rotate and diffuse, so neither method of crystal delivery can pre-align the crystal with the X-ray beam. Hence, computational alignment of the diffraction images is non-trivial: every image has to be indexed separately and many images contain just too few reflections to allow robust determination of cell constants. In case several indexing matrices are allowed by the space group, all of them need to be tried to group images into sets with consistent indexing. Scaling of these grouped images into a complete dataset is complicated by several factors. First, the overall mean intensities on the images are different depending on crystal volume and whether the crystal was hit by the X-ray beam full-on or just grazed by it. Sample preparation therefore aims at a very narrow distribution of crystal sizes. Producing tiny droplets that contain little solvent besides the crystal may reduce the grazing probability and also reduces background scattering by the droplet solution. Second, if the cell dimensions of the crystals are not isomorphous, scaling will lead to large R-values, and the information content in the dataset is reduced. Many diffraction images that do not scale well with others are discarded during the process to arrive at a complete dataset, a computationally and energetically (for electricity costs and CO_2 production) very expensive undertaking. Until now, most diffraction data from serial crystallography led to lower-quality models compared to models from a single-crystal data.

XFELs are mainly used for serial crystallography experiments where synchrotrons are insufficient, *i.e.* nanometer-sized crystals and rapid chemical reactions. The light intensity generated

by the currently strongest XFELs (European XFEL, SwissFEL, LCLS) is about 10^9-fold higher than that from a conventional third-generation synchrotron, enabling radiation damage-free time-resolved and time-sensitive studies at room temperature. This includes structures of reaction intermediates in enzymatic reactions and metalloproteins in their oxidized form. Enzymatically active nanocrystals can be soaked with substrate and when using a conveyor belt for crystal delivery, a time shift can be set until the XFEL illuminates the crystal. Thus, intermediates along the reaction coordinate may be populated and their structure elucidated.

X-rays emitted from a wiggler have a uniform spatial distribution, which means the radiation they emit is incoherent: there is no phase relation between the X-rays. The situation is the same for an electron bunch entering an undulator. However, once in the undulator, the electrons start to interact with the electromagnetic radiation they have produced themselves. The X-rays emitted travel faster than electrons, overtaking electrons ahead in the bunch and interacting with them on the way. The interaction accelerates some electrons and slows down others, depending on the phase relation between the electron and the X-rays. As a result, the electrons gradually drift into a series of thin discs, termed *micro-bunches*. The micro-bunches contain about 10^8–10^9 electrons each, are separated by a distance equal to one wavelength, and start to emit coherent (same phase) X-rays that reinforce themselves by constructive interference. A runaway effect ensues: the increase in light intensity further increases the sharpness of the micro-bunches by the same kind of X-ray/electron interaction as above, which in turn increases coherence. The effect is known as *self-amplified spontaneous emission* (SASE; Figure 24.31). The separation of electrons into micro-bunches in the undulator takes some time, so XFELs require very long undulators of > 100 m. At the end of the undulator, the electrons are magnetically dumped into a cavity so they will not damage the X-ray detector and the light is transmitted onto the sample.

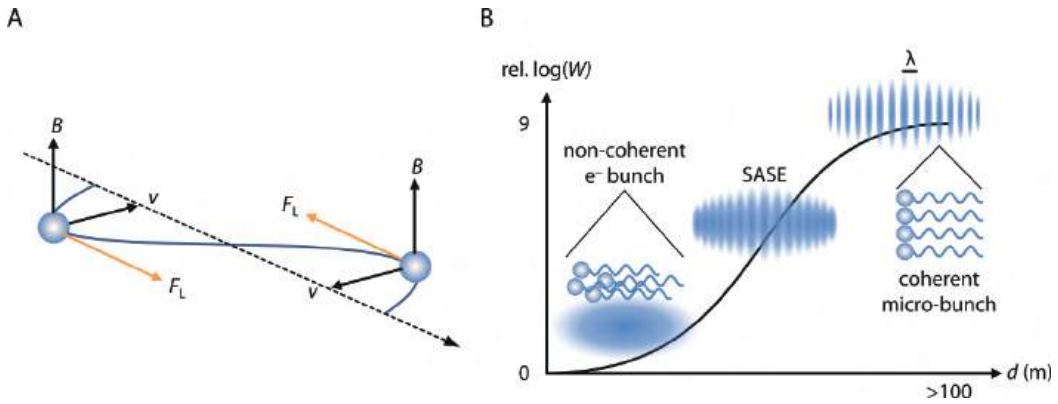


FIGURE 24.31: Self-amplified spontaneous emission. A: The sinusoidally moving electrons (blue) in an undulator emit X-rays. The magnetic field component *B* of the X-rays couples with the transverse component *v* of the electron movement. The resulting Lorentz force F_L acts on the electrons along the direction of their travel (dashed line), slowing those that are too fast and accelerating those that are too slow. The bunching leads to constructive interference of the X-rays, which in turn increases the coupling. B: Along the length *d* of the undulator that can reach a few hundred meters, the increasing bunching boosts the output power *W* of the X-rays by about nine orders of magnitude. The final coherent electron bunch encompasses a series of micro-bunches separated by a distance equal to λ. This fine structure of the X-ray pulse can be used in time-resolved experiments on the femtosecond scale.

XFELs can be tuned easily by changing either the electron beam energy of the storage ring feeding them or the magnetic field strength of the long undulator. The light has the high spatial and temporal coherence that is characteristic of lasers, but is limited to a single pulse. The spatial coherence allows strong focusing of the laser light into a tiny spot, while the temporal coherence provides a sequence of extremely short (fs) and intense X-ray pulses whose timing can be changed. As of 2024, there are slightly over 400 crystal structures reported in the PDB with the term "XFEL", of which less than 170 have different sequences. The resolution of these structures ranges from 1.0 Å to 5.2 Å with the objects imaged from simple trans-membrane helices to 1 MDa photosystems.

The next step in the development of XFELs would be to determine structures of single molecules. This would require recording of the scattering patterns of a series of single molecules in random orientations, aligning the 2D patterns, assembling them into a three-dimensional intensity distribution, and phasing the structure. While theoretically possible, this has not yet been achieved mainly due to too low scattering signal from a single molecule. At present, structures determined using an XFEL source still require crystals.

24.7 NEUTRON CRYSTALLOGRAPHY

If neutrons instead of X-rays are used in a diffraction experiment, a neutron crystal structure can be determined in the same way as an X-ray crystal structure. While X-rays are scattered at electrons, neutrons are scattered at nuclei. Currently available neutron beams are several orders of magnitude less intense than X-ray beams, and neutron detectors are not as sensitive as X-ray detectors. Therefore, the major requirement for a neutron crystal structure is large crystals of at least 1 mm in each direction. However, the advantage is that with neutron crystallography hydrogen atoms can be visualized, something that is difficult to achieve with X-rays. Neutron crystallography is not used for *de novo* structure determination (this is done with X-rays), but is used to study hydrogen atoms, particularly to elucidate enzymatic mechanisms.

While the result of X-ray crystallography is an electron density map, neutron crystallography returns a *nuclear density map*, *i.e.* the distribution of atomic nuclei within a molecule. Hydrogen isotopes are not normally visible in electron density maps because they contain only a single electron, and that electron is shifted toward the heavier atom in the X–D bond. In contrast, hydrogen and deuterium show up very prominently in nuclear density maps as negative and positive peaks, respectively. This arises since the scattering lengths of hydrogen (–3.7 fm) and deuterium (+6.7 fm) are comparable in magnitude to the scattering lengths of other atoms (see Table 23.2) and hence give a clear signal. Deuterium has a scattering length comparable to carbon, and hydrogen is unique in having a scattering length of negative sign, making it particularly easy to distinguish from other nuclei by its negative nuclear density. The large signal of hydrogen and deuterium in nuclear maps is used to study the protonation states of proteins: crystals are soaked in buffer containing H_2O, D_2O, or a mixture, then flash-cooled so that no further H/D exchange can take place, and a series of neutron crystal structures is determined by molecular replacement using the X-ray crystal structure as a search model. Enzyme reactions with shifts in hydrogen atoms are particularly suited for neutron crystallography. For example, the location of the proton in the catalytic Ser-His-Asp triad of serine proteases was detected by neutron crystallography to be at the His, rather than at the Asp. In D-xylose isomerase, a metalloenzyme catalyzing the isomerization of aldoses to ketoses, four reaction intermediates have been detected using neutron crystallography.

24.8 ELECTRON CRYSTALLOGRAPHY

The low signal in electron microscopy (EM) images is one of the main reasons for the comparatively low resolution of EM structures. To improve the signal in cryo-EM, images are taken from a large number of molecules, aligned, and averaged (Chapter 26). Alternatively, pre-aligned molecules can be studied by electron crystallography. Aligning and averaging images of single molecules from different orientations is conceptually identical to pre-ordering the molecules first in a crystal lattice and then imaging them. As crystals are arrays of very well-aligned molecules, the electron dose used in electron crystallography is smaller than that needed for single-particle cryo-EM. The wavelength of the incident electrons is small compared to atomic distances, so electron crystallography has the potential to produce high-resolution information on atomic positions. Electron crystallography combines advantages from microscopy and crystallography by collecting both images and electron diffraction patterns from crystals. From the images, the phases of the object may be retrieved, and the crystal amplifies the signal in the diffraction pattern. Due to the strong interaction of electrons with

matter, the crystals used for electron crystallography need to be thin, ideally less than 100 nm. Two-dimensional sheets or helical crystals produced by fibers are commonly used, but three-dimensional micro-crystals also work. Membrane proteins, which are naturally dissolved in two-dimensional lipid bilayers *in vivo*, can sometimes be made to associate into two-dimensional crystals. Two-dimensional, or thin three-dimensional crystals of soluble proteins may also be obtained by ordered assembly on a helper surface, such as mica or a lipid monolayer. The diffraction by these sheet-like crystals is continuous in the direction perpendicular to the crystal plane. The resulting lattice lines are sampled by collecting data at different tilt angles (±70°). Just as with cryo-EM imaging and electron tomography, there will almost always be a missing part of data that affects the quality of the resulting electron density map. In addition, the thin crystals are prone to bending during sample preparation, an effect that can be computationally corrected if it is small (*un-bending*).

As of mid-2024, the PDB holds ~240 structures determined by electron crystallography, with 72 unique structures at the 95% sequence similarity level. Among these are amyloid-forming peptides, soluble proteins (lysozyme, catalase, proteinase K, trypsin), muscle and fiber proteins (actin, myosin, tubulin), and membrane proteins (GPCRs, bacteriorhodopsin, K^+ channel, Ca^{2+} ATPase, prostaglandin E synthase). The first structure of bacteriorhodopsin had a resolution of only 0.35 nm, showing the orientation of α-helices. More details were visible in the structure of the membrane protein aquaporin AQP0, which has been determined from double-layered, two-dimensional crystals at a resolution of 0.19 nm, including the lipid bilayer surrounding the AQP0 tetramers and the water molecules within the pores (Figure 24.32).

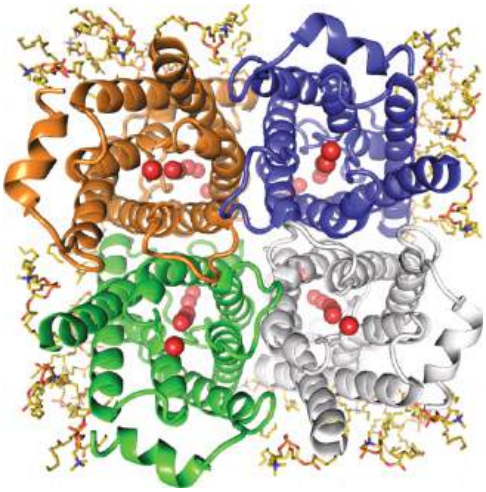


FIGURE 24.32: Structure of aquaporin. The membrane protein aquaporin forms homotetramers of four water-conducting pores. At the center of each channel there are three water molecules (red spheres) that are too far apart to form hydrogen bonds. The lipids (yellow stick models) packing around the tetramer adopt preferred conformations. Gonen et al. (2005). Nature 438(7068): 633–638.

QUESTIONS

24.1 Which parameters can (and should) be varied in a crystallization experiment?

24.2 Name the two important steps during crystallization and explain the rationale behind seeding.

24.3 Why is it so difficult to crystallize macromolecules, and especially membrane proteins?

24.4 What are the main side-reactions that compete with crystallization?

24.5 What are options to render a flexible protein more likely to crystallize?

24.6 You can crystallize a nucleotide-binding protein from 2 M $(NH_4)_2SO_4$ but fail to co-crystallize a protein-nucleotide complex under these conditions. What could be the reason?

24.7 Moseley's law is an empirical law that relates the energy of the electronic transitions to the atomic number Z: for the K_α transitions, Moseley's law is

$$\lambda = \frac{121568.5}{(Z-1)^2}\, pm$$

Calculate the expected wavelengths for Cu and Ni.

24.8 The K_β transitions for Cu and Ni are at 139 pm and 150 pm, respectively. Why can a thin Ni-foil be used to remove the K_β radiation from the spectrum generated by a Cu anode?

24.9 A synchrotron storage ring has electrons moving at 4.5 GeV energy. At which velocity in percent relative to the speed of light do the electrons travel and how does their relativistic mass compare to a resting neutron? An experimental station is placed at d = 10 m tangential to the ring. What is the maximum width w of the radiation cone at this position and how does it compare with the typical dimensions of a crystal?

24.10 Calculate the number of unit cells in a cube-shaped crystal of 0.1 mm edges for unit cell dimensions $a = b = c$ = 10 nm. For some XFEL experiments, crystals of only 100 unit cells along each edge are sufficient. What edge would a cube-shaped crystal with this number of unit cells have?

24.11 A protein of molecular mass 21 kDa has been crystallized in tetragonal space group $P4_12_12$ (8 symmetry operators) with cell dimensions $a = b$ = 98.5 Å, c = 104.2 Å, $\alpha = \beta = \gamma = 90°$. How many molecules are possibly in the asymmetric unit and what is the solvent content in the most likely case? The solvent content is related to the Matthews parameter V_M by the following rule of thumb:

$$\text{Solvent} = 1 - \frac{1.23}{V_M}$$

24.12 A protein of 89 amino acid residues has been crystallized in space group $P2_12_12_1$ (4 symmetry operators) with cell dimensions of a = 28.4 Å, b = 36.3 Å, and c = 65.7 Å. What is the likely number of molecules per asymmetric unit and what is the solvent content of this crystal? The average molecular mass for an amino acid residue is 112.5 Da.

24.13 What is the lowest resolution data one can collect from a cubic crystal of cell edges 5 nm? What is the diffraction angle corresponding to this resolution when using X-rays of 0.1 nm wavelength? What is the maximum radius a circular beam stop placed 20 mm from the crystal can have in order to let this reflection pass on to the detector? How far does a beam stop of 1 mm diameter need to be placed away from the crystal to allow recording of this reflection? How far away from the crystal does a circular detector of 345 mm diameter need to be in order to collect data to a high-resolution limit of 0.2 nm?

24.14 Name at least two advantages of collecting X-ray diffraction data at 100 K compared to room temperature.

24.15 Why are only 65 out of the total 230 space groups accessible to most biological molecules?

24.16 Many space groups have symmetry operators that lead to "reflections" of zero intensity, termed *systematically absent reflections*. Using the structure factor formula

$$\boldsymbol{F}_{\mathbf{hkl}} = \left|F_{\mathrm{hkl}}\right| \cdot e^{i\phi_{\mathrm{hkl}}} = \sum_{n=1}^{N} f_{\mathrm{n},\theta_{\mathrm{hkl}}} \cdot e^{2\pi i \cdot \left(hx_{\mathrm{n}} + ky_{\mathrm{n}} + lz_{\mathrm{n}}\right)}$$

find the systematically absent reflections along the *k*-axis, *i.e.* the (0*k*0) reflections, for space group $P2_1$. This space group has the symmetry operators (*x*,*y*,*z*) and (–*x*,*y*+½,–*z*). The symmetry operators denote equivalent positions in the crystal and the diffraction from these positions is additive.

24.17 Using the same approach as in the previous question, show that I-centered lattices give rise to only half the number of reflections compared to a primitive lattice of the same cell dimensions. For each position (*x*,*y*,*z*), I-centered space groups have an equivalent position at (*x*+½,*y*+½,*z*+½).

24.18 Why are heavy atom derivatives for nucleic acids more easily obtained using heavy cations (*e.g.* Ba^{2+}) than heavy anions (*e.g.* I^-).

24.19 Why is it useful that Friedel's law is not always fulfilled?

24.20 Can one distinguish between the two 180° rotated orientations of the imidazole ring of a buried His side-chain solely based on electron density?

24.21 Assign the suitability for (1) restraints during refinement and (2) validation of the model obtained for: bonds, 1,3-angles, torsion angles, electron density real-space correlation coefficient, side-chain rotamers, planarity of aromatic side-chains and peptide bonds, hydrogen bonds, geometric definition of ligands.

REFERENCES

Bourgeois, D. and Royant, A. (2005) Advances in kinetic protein crystallography. *Curr Opin Struct Biol* **15**(5): 538–547.
· review on the Laue method for kinetic crystallography and its complementation with other spectroscopic techniques

Meents, A., Gutmann, S., Wagner, A. and Schulze-Briese, C. (2010) Origin and temperature dependence of radiation damage in biological samples at cryogenic temperatures. *Proc Natl Acad Sci* **107**: 1094–1099.
· a review on neutron crystallography

Brünger, A. T. (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**: 472–475.
· introduces the free *R*-factor for global model validation

Derewenda, Z. S. (2010) Application of protein engineering to enhance crystallizability and improve crystal properties. *Acta Crystallogr D Biol Crystallogr* **66**(Pt 5): 604–615.
· overview of protein-engineering methods designed to enhance crystallizability

Glusker, J. P., Carrell, H. L., Kovalevsky, A. Y., Hanson, L., Fisher, S. Z., …, Langan, P. (2010) Using neutron protein crystallography to understand enzyme mechanisms. *Acta Cryst D* **66**(Pt 11): 1257–1261.

Owen, R. L., Rudino-Pinera, E. and Garman, E. F. (2006) Experimental determination of the radiation dose limit for cryocooled protein crystals. *Proc Natl Acad Sci* **103**: 4912–4917.

Holton, J. M. (2009) A beginner's guide to radiation damage. *J Synch Radiat* **16**(Pt 2): 133–142.
· review on radiation damage

Kern, J., Yachandra, V. K. and Yano, J. (2015) Metalloprotein structures at ambient conditions and in real-time: Biological crystallography and spectroscopy using X-ray free electron lasers. *Curr Opin Struct Biol* **34**: 87–98.
· XFELs can help avoid the radiation damage typically observed at the metal centers of metalloproteins

Matthews, B. W. (1968) Solvent content of protein crystals. *J Mol Biol* **33**: 491–497.
· introduction of the concept that V_M and solvent content are distributed over a rather narrow range in protein crystals

McCoy, A. J., Sammito, M. D. and Read, R. J. (2022) Implications of AlphaFold2 for crystallographic phasing by molecular replacement. *Acta Cryst D* **78**(Pt 1): 1–13.
· summarizes possibilities and pitfalls for the use of AlphaFold2 models in molecular replacement

Murshudov, G. N., Vagin, A. A. and Dodson, E. J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst D* **53**(Pt 3): 240–255.
· the application of the maximum likelihood target for crystallographic refinement in modern software is described

Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. and Hajdu, J. (2000) Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**(6797): 752–757.
· suggests that single-molecule "crystallography" is in principle possible

Pannu, N. S. and Read, R. J. (1996) Improved structure refinement through maximum likelihood. *Acta Cryst A* **52**: 659–668.

Blakeley, M. P., Hasnain, S. S. and Antonyuk, S. V. (2015) Sub-atomic resolution X-ray crystallography and neutron crystallography: Promise, challenges and potential. *IUCrJ* **2**(Pt 4): 464–474.
· neutron crystallography reveals intermediates for aldose/ketose isomerization

Read, R. J. (1986) Improved fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst A* **42**: 140–149.

Redecke, L., Nass, K., DePonte, D. P., White, T. A., Rehders, D., …, Chapman, H. N. (2013) Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser. *Science* **339**(6116): 227–230.
· the structure of the *Trypanosoma brucei* cysteine protease cathepsin B was determined using an XFEL from micro-crystals that spontaneously grew *in vivo*

Rupp, B. (2010) *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. New York, Garland Science.
· a comprehensive and up-to-date treatment of crystallography for the biological sciences

Tickle, I. J., Laskowski, R. A. and Moss, D. S. (1998) R_{free} and the R_{free} ratio. I. Derivation of expected values of cross-validation residuals used in macromolecular least-squares refinement. *Acta Cryst D* **54**: 547–557.

Tickle, I. J., Laskowski, R. A. and Moss, D. S. (2000) R_{free} and the R_{free} ratio. II. Calculation of the expected values and variances of cross-validation statistics in macromolecular least-squares refinement. *Acta Cryst D* **56**(Pt 4): 442–450.
· describes the relation of the crystallographic and free *R*-factors

Warkentin, M. and Thorne, R. E. (2007) A general method for hyperquenching protein crystals. *J Struct Funct Genomics* **8**(4): 141–144.
· to reduce the concentration of cryo-protectant, a simple technique termed *hyperquenching* may be used

Winn, M. D., Isupov, M. N. and Murshudov, G. N. (2001) Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Cryst D* **57**(Part 1): 122–133.
· TLS refinement for anisotropic B-value modeling at low resolution

ONLINE RESOURCES

kinemage.biochem.duke.edu/teaching/BCH681/2013BCH681/elasticScattering
· Elastic scattering explained with relation to X-ray crystallography.

www.xtal.iqfr.csic.es/Cristalografia/index-en.html
· Comprehensive overview of crystallography with many figures, animations, and historic links.

www.doitpoms.ac.uk/tlplib/index.php
· The subsections on "Crystallography", "Reciprocal Lattice", and "X-ray Diffraction Techniques" contain many instructive animations.

www.ysbl.york.ac.uk/~cowtan
· Kevin Cowtan's tutorials on structure factors and crystallographic Fourier transforms.

25

Fluorescence imaging and microscopy

In the previous Chapters, we have discussed methods that allow us to measure physical parameters for ensembles of isolated molecules (Chapter 21) or methods that provide structural information on biomolecules (Chapters 22, 23, 24). Microscopes are devices that generate magnified images of small objects such as cells and sub-cellular structures by using optical principles (light microscopy), streams of charged particles (particle microscopy; Chapter 26), or scanning probes (scanning probe microscopy; Chapter 27).

In fluorescence microscopy, a type of *light microscopy*, fluorescence emitted by molecules in the sample is used to generate images of tissues, cells, and sub-cellular structures. By providing spatial resolution, fluorescence microscopy allows localization of fluorescent(ly labeled) biomolecules within the cell. The cellular localization of biomolecules provides hints as to in which processes they are involved and what their particular function might be. Fluorescence microscopy can also be used to determine fluorescence parameters for single molecules, either in solution or immobilized on surfaces, which reveals the distribution of the respective parameter over all molecules (Section 25.7).

In a fluorescence microscope, the sample is illuminated by a light source, and the fluorescence from the object of interest is detected as a function of position. Depending on the illumination geometry, a small spot within the sample is illuminated, while the rest remains in the dark, or a larger area is illuminated with uniform light intensity. The fluorescence emitted from the illuminated spot or area is collected by a microscope objective. Excitation light is rejected by a filter, and the emitted fluorescence light is passed on to a point or area detector. Typical point detectors are sensitive photomultipliers or avalanche photodiodes. To generate an image of the sample, the fluorescence of the sample is measured as a function of position, point by point. The intensity is detected by a point detector, one pixel after the other. Alternatively, and more commonly, the fluorescence intensity from an area is imaged onto an array detector, such as a CCD camera.

25.1 OPTICAL PRINCIPLES OF MICROSCOPY

Before we compare different types of fluorescence microscopy in more detail, we will first review the very basic principles of optics. Microscope objectives are nothing other than optical lenses, and to

appreciate the differences between microscope types, we need to understand how these lenses focus and collect light.

25.1.1 Focusing and collecting light by optical lenses

Light is collimated or focused by collecting lenses with convex surfaces. When illuminated by a parallel light bundle, the light is refracted at the air-glass interface when it enters the lens, and at the glass-air interface when it exits the lens. This refraction at the entry and exit interface leads to *focusing* of the light bundle into a small spot in the focal plane of the lens (Figure 25.1). The distance between the lens and the focal spot is called the *focal length f*. Focusing lenses are used to focus excitation light into small spots within the microscope sample.

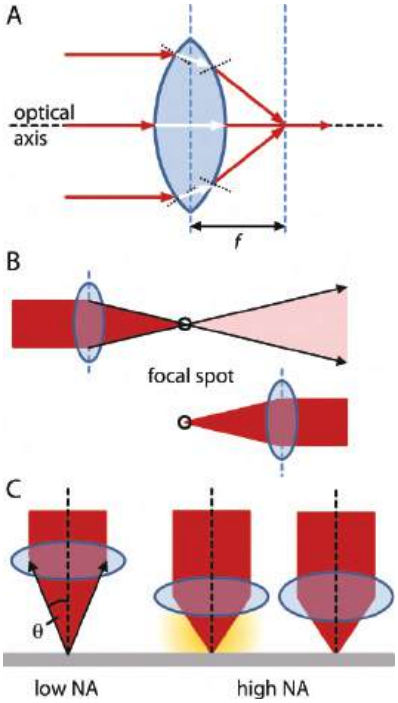


FIGURE 25.1: Focusing lenses and principles of optics. A: A focusing lens. Only the outer and central rays of the parallel light bundle are depicted (red arrows). At the air-glass interface, the rays are refracted toward the optical axis (black dotted lines). The light path within the lens is indicated by white arrows. At the distal side of the lens, the exiting rays are again refracted toward the optical axis. As a consequence, all rays meet in the same spot on the optical axis in the focal plane. The distance between the center of the lens and the focal plane is called the focal distance *f*. B: Focusing and collimation of light. Top: Parallel light (red) is focused into a focal spot by a focusing lens. After the focal spot, the light bundle diverges again (light red). Bottom: Light from a point source in the focal plane of a collimating lens is collected by the lens and converted into a parallel light bundle. C: Numerical aperture. The numerical aperture NA depends on the highest angle θ of rays that are collected by a lens. The NA can be increased by placing a lens with a shorter focal length closer to the light source (right), or by using an immersion oil (yellow) with a higher refractive index than air (center).

Optical light paths are reversible, meaning that light emitted from a point source in the focal plane of a focusing lens will end up as a parallel light bundle on the opposite side of the lens, which is called *collimation*. This is the arrangement that allows us to use lenses to collect fluorescence light emitted from a certain spot within our microscope sample. The *light collection efficiency* of a lens is described by the *numerical aperture* NA (Figure 25.1). The numerical aperture is defined as

$$NA = n \sin\theta \qquad \text{eq. 25.1}$$

n is the refractive index of the medium between the point source and the lens, and θ is the half-angle of the outer rays of the collimated light with the optical axis. The larger θ, the more light is collected by the lens. The angle θ depends on the focal length f of the lens and its radius r according to

$$\sin\theta = \frac{r}{f} \qquad \text{eq. 25.2}$$

A higher collection efficiency, *i.e.* higher NA, is achieved either by using a lens with a shorter focal length f and/or by a larger radius (eq. 25.2), or by using an immersion liquid with a refractive index n higher than air, such as water or oil (eq. 25.1).

Lenses can be combined to manipulate light beams in various ways. One example is the combination of two collecting lenses in a *telescopic lens system* to widen or narrow a light beam (Figure 25.2). A parallel light bundle of radius r_1 is focused by a first lens with a focal length f_1. A second lens is placed such that its distance from the focus equals its focal length f_2. The light diverges as it propagates from the focal spot, and is collimated and converted into a parallel light bundle after passing the second lens. The radius r_2 of the light bundle is determined by the ratio of the focal lengths:

$$\frac{r_1}{r_2} = \frac{f_1}{f_2} \qquad \text{eq. 25.3}$$

According to eq. 25.3, the beam is widened for $f_2 > f_1$, and the beam is compressed for $f_1 > f_2$.

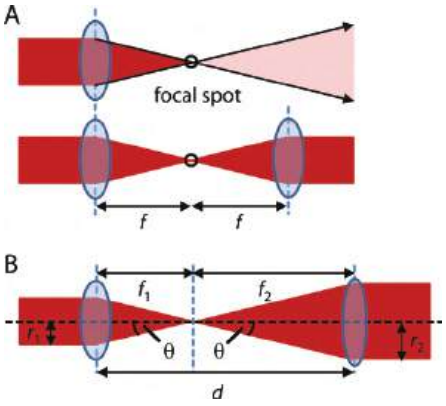


FIGURE 25.2: Shaping beams by telescopic lens systems. A: A parallel light bundle is focused by a lens and diverges after the focal spot (top). A second lens placed at its focal distance f from the focal point collects the light from the focal point and converts it back into a parallel light bundle (collimation, bottom). B: Two lenses with different focal lengths f_1 and f_2 are placed at a distance $d = f_1 + f_2$, such that their focal points coincide. Such a telescopic lens system increases or decreases the radius of the parallel beam. The ratio of the radii r_1/r_2 depends on the ratio of the focal lengths f_1/f_2 (eq. 25.3).

Special lens shapes create special effects. We have seen before that a convex lens focuses light into a focal spot. A *cylindrical lens*, on the other hand, is only curved in one dimension (Figure 25.3). Therefore, it focuses light only in or from the plane perpendicular to the cylinder axis. As a result, it generates a *focal line* instead of a focal point (Figure 25.3). We will see later that this property can be exploited to illuminate samples within defined areas (see Box 25.2).

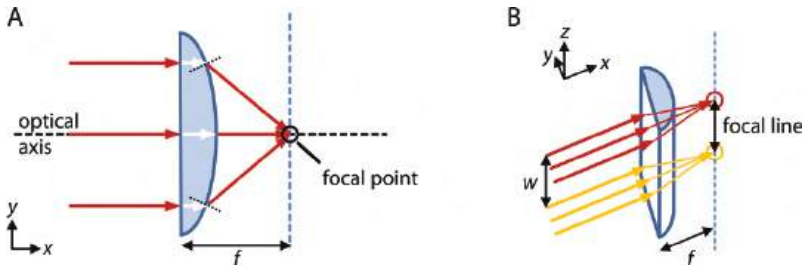


FIGURE 25.3: A cylindrical lens focuses light only in one dimension. A: Top view: A cylindrical lens is illuminated with a parallel light bundle that impinges on the planar side at a right angle and is not refracted. Light exiting the lens on the convex side is refracted and focused into a focal point. B: Light hitting the lens in different positions along the *z*-axis (red, orange) is focused into different points in the focal plane. A light bundle of the width *w* in the *z*-direction leads to a focal line extending from the focal point of one limiting beam to the focal point of the second limiting beam (red and orange circles).

25.1.2 Microscopes: How to achieve magnification with optical lenses

The size an object appears depends on the distance d of the eye from the object and the angle ε at which the outer rays reach the eye (Figure 25.4). A *magnification* of 1 is defined by the angle ε_0 at which we see an object that is located at a distance d = 25 cm from the eye. At larger distances ($\varepsilon < \varepsilon_0$), the same object appears smaller, at shorter distances ($\varepsilon > \varepsilon_0$), it appears larger. The magnification m is defined by the ratio of the corresponding angles

$$m = \frac{\varepsilon}{\varepsilon_0} \qquad \text{eq. 25.4}$$

In principle, we can achieve increasing magnifications by moving our eyes closer and closer to the object. However, there is a limit to this approach. The eye cannot accommodate distances below ca. 10 cm. To reach higher magnifications, we therefore need optical instruments such as a magnifying glass or a microscope. A minimal *microscope* consists of two focusing lenses (Figure 25.4): the first lens, the objective, collects light from the object O and generates an intermediate image I that is observed through a second lens, the ocular.

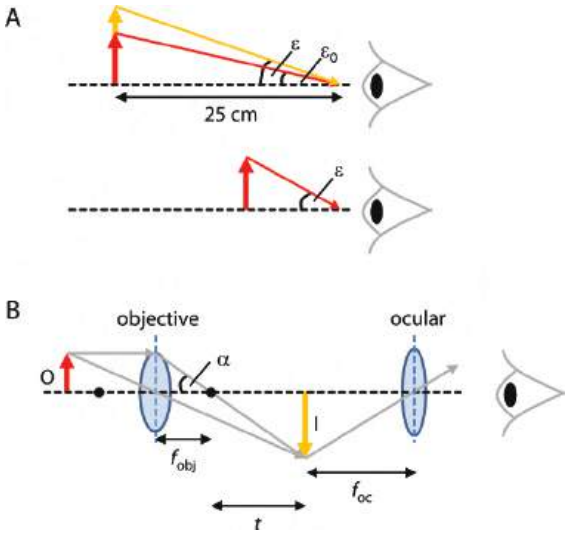


FIGURE 25.4: Microscopes and magnification. A: The size an object appears depends on the angle ε. A magnification of 1 is defined to correspond to an object at a distance of 25 cm from the eye, giving the angle ε_0 (top, red). Larger objects lead to larger angles ε (orange), smaller objects lead to smaller angles. An object of the same size, at a shorter distance to the eye (bottom), is perceived larger. The magnification m is defined by the ratio $\varepsilon/\varepsilon_0$. B: A microscope consists of an objective lens and an ocular lens. The objective generates an image I of the object that is viewed through the ocular. f_{oc}, f_{obj}: focal lengths of ocular and objective, t: tube length.

The ratio of the sizes of object O and intermediate image I defines the magnification m_{obj} of the microscope objective:

$$m_{obj} = \frac{I}{O} \qquad \text{eq. 25.5}$$

According to Figure 25.4, we can relate this ratio to the focal length f_{obj} of the objective and the distance between the focus of the objective and the objective image, the tube length t. From the geometric relationship

$$\tan\alpha = \frac{O}{f_{obj}} = \frac{I}{t} \qquad \text{eq. 25.6}$$

we obtain

$$m_{obj} = \frac{I}{O} = \frac{t}{f_{obj}} \qquad \text{eq. 25.7}$$

The ocular also contributes to the magnification. The magnification m_{oc} is given by the ratio of the intermediate image I and the image I' perceived by the eye:

$$m_{oc} = \frac{I'}{I} \qquad \text{eq. 25.8}$$

The size of the perceived image I' depends on the distance d of the eye from the ocular and on the angle α according to

$$\tan\alpha = \frac{I'}{d} = \frac{I}{f_{oc}} \qquad \text{eq. 25.9}$$

By rearranging eq. 25.9, we obtain the magnification m_{oc} as

$$m_{oc} = \frac{d}{f_{oc}} \qquad \text{eq. 25.10}$$

The overall magnification m of a microscope is the product of objective and ocular contributions:

$$m = m_{oc} \cdot m_{obj} = \frac{d}{f_{oc}} \cdot \frac{t}{f_{obj}} \qquad \text{eq. 25.11}$$

In advanced microscopes, the image is not observed by eye, but is imaged onto a detector that determines the light intensity at each spot of the image.

25.1.3 The diffraction limit of optical resolution

Strictly speaking, a lens can only focus light into a focal point in the absence of diffraction. However, when a lens is illuminated, light is always diffracted. The diffraction widens the focal spot to an intensity distribution in the focal plane that is centrosymmetric around the optical axis, the *Fraunhofer diffraction pattern*. The Fraunhofer patterns of a uniformly illuminated centric aperture such as a lens consists of a bright circular region in the center of the focus, surrounded by a set of concentric rings of lower intensity (Figure 25.5). Such a pattern is also called a *diffraction-limited focus*. The diameter of the central bright area, the *Airy disc*, is defined by the position of the first minimum of the diffraction pattern. The angle θ_{min} at which the first minimum is observed depends on the wavelength λ and the diameter d of the lens:

$$\sin\theta_{\min} = 1.22\frac{\lambda}{d} \qquad \text{eq. 25.12}$$

The number 1.22 takes into account the circular shape of the lens aperture and the diffraction pattern. For small angles, we can approximate

$$\theta_{\min} \approx 1.22\frac{\lambda}{d} \qquad \text{eq. 25.13}$$

The diffraction-limited focus is the smallest possible focus we can achieve when focusing light by optical lenses. Conversely, when we collect light from a microscope sample and focus it onto a detector, a point source of light within the sample is detected as the same centrosymmetric intensity distribution, the *point-spread function* (PSF). The PSF is the intensity pattern obtained by imaging a point source of light through a microscope. Typically, the rings surrounding the Airy disc are not detected because of their low intensity. Hence, a point source of light within the sample is detected as a finite circular spot. The intensity distribution as a function of the *x*- and *y*-coordinates can be described by Gaussian distributions. This detected spot is called the *molecule detection function* (MDF; Figure 25.5).

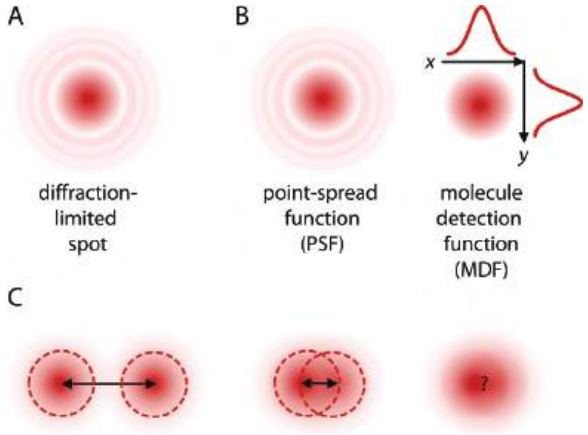


FIGURE 25.5: Diffraction, point-spread function, molecule detection function, and optical resolution. A: Due to diffraction, light cannot be focused into an infinitely small spot. The intensity distribution of a diffraction-limited spot is centrosymmetric around the optical axis. The central disc of high intensity, the Airy disc, is surrounded by concentric rings of decreasing intensity. B: A point source of light is imaged by a lens according to the point-spread function (PSF), with the same intensity pattern as a diffraction-limited spot (left). The intensity of the surrounding rings is often too low to be detected. The diffraction-limited image from a point source of light is therefore a centrosymmetric intensity distribution corresponding to the Airy disc (right). This intensity distribution is called the molecule detection function (MDF). C: The size of the Airy disc determines the resolution of a microscope. Two point sources of light can be detected as separate if the distance between their centers is larger than the radius of the Airy disc.

The *resolution* of a microscope is a measure of the minimal distance *d* between two points that are still detected as distinct. Ernst Abbe found the relationship

$$d < \frac{\lambda}{2n\sin\theta} = \frac{\lambda}{2NA} \qquad \text{eq. 25.14}$$

for the resolution of light microscopes. For air objectives (n = 1), we obtain the well-known limiting resolution

$$d < \frac{\lambda}{2} \qquad \text{eq. 25.15}$$

Microscopy using light in the visible range (400–800 nm) therefore cannot achieve resolutions below 200–400 nm.

For microscopy of light-emitting samples, the resolution is determined by the size of the Airy disc of the PSF or MDF: two point sources that are far apart in the object are detected as two individual diffraction patterns (Figure 25.5). When the points come closer, the MDFs will start to overlap. The shortest distance at which the points can still be distinguished is reached when the first minimum of the pattern from one point source coincides with the central maximum from the other point source. In other words, the distance between the centers of the Airy discs must be larger than the disc diameter. This condition is called the *Rayleigh criterion*. The diameter d of the Airy disc can be approximated as

$$d \approx 0.61 \frac{\lambda}{n \sin\theta} = 0.61 \frac{\lambda}{NA} \qquad \text{eq. 25.16}$$

For light in the visible range (400–800 nm) and standard oil immersion objectives with numerical apertures of 1.4, the diameter of the Airy disc and thus the diffraction-limited resolution is ca. 200–400 nm (in the xy-plane).

Diffraction only becomes relevant when light propagates over long distances. Near-field methods avoid blurring because of diffraction by obtaining the signal near the emitting molecule. In *near-field scanning optical microscopy*, a resolution of 20–50 nm is achieved by scanning a sharp tip across the surface to probe the sample. However, because of their restriction near to surfaces, these methods are of limited value for high-resolution microscopy on biological specimen. We will see later that far-field microscopy can also be taken to higher resolutions beyond the diffraction limit by introducing modifications that compensate for diffraction effects (see Section 25.8).

25.2 WIDE-FIELD FLUORESCENCE MICROSCOPY

Fluorescence microscopy uses fluorescence excitation and detects fluorescence emitted from the sample as a function of position. In epi-fluorescence microscopy, the objective is used to illuminate the sample with light and to collect the fluorescence emission. Standard fluorescence microscopes illuminate the sample from above, whereas *inverted microscopes* illuminate the sample from below, and allow easy access for additional manipulation from above. In *wide-field microscopy*, a parallel light bundle is focused into the back-focal plane of the objective (Figure 25.6). As a result, the light is again parallel on the distal side of the objective lens, allowing homogenous illumination of a defined area on the microscope slide. Fluorophores within this illuminated area are excited and emit fluorescence in all three directions of space. The fraction emitted toward the objective is collected by the objective lens, passed through a filter (Box 25.1) to reject any scattered excitation light, and imaged onto a detector.

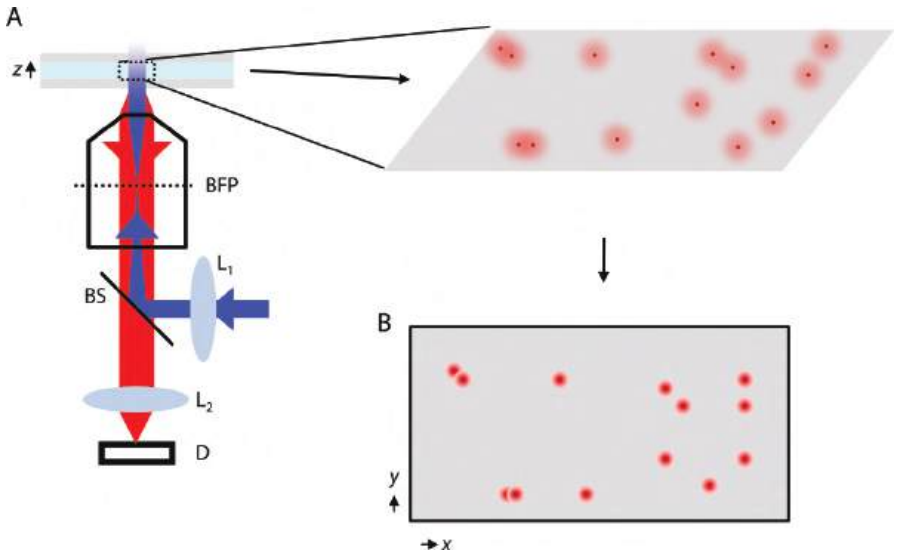


FIGURE 25.6: Wide-field microscopy. A: Excitation light (blue) is focused into the back-focal plane (BFP) of the objective by lens L_1. After passing the objective lens, a parallel light bundle is generated that illuminates an area of the sample (inset). Fluorescent light emitted from the sample (red) is collimated by the objective, passes the beam splitter (BS), and is focused onto a detector D by a second lens (L_2). B: On the detector, each fluorescent emitter in the sample generates a point-spread function. In a wide-field microscope, all fluorophores in the light path along the *z*-axis are illuminated and detected. The image is a projection of all these emitters into one plane.

BOX 25.1: OPTICAL FILTERS.

Optical filters enable the selection of defined spectral regions of light, and are used to separate excitation light from red-shifted emission or to separate fluorescence from different emitters with different fluorescence wavelengths (Figure 25.7). A longpass (LP) filter transmits light above a cut-off wavelength and rejects all wavelengths below. These filters are named LP, followed by a number that corresponds to the wavelength at which the light is transmitted (cut-off wavelength). Long-pass filters with a very sharp transition between transparency and absorption are called edge filters. A short-pass filter (SP) is transparent for light below a threshold wavelength, but rejects light above this wavelength. A bandpass filter (BP) is transparent for a certain range of wavelengths, but rejects all wavelengths above and below. These filters are named BP, followed by the lower and higher cut-off wavelength. A beam splitter or dichroic mirror (DM) is an optical element that is transparent for light above a certain wavelength. Light of lower wavelengths is reflected.

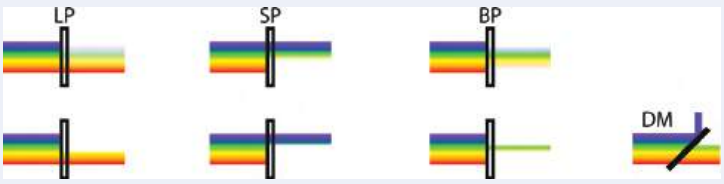


FIGURE 25.7: Optical filters. LP longpass filter, SP short path filter, BP bandpass filter, DM dichroic mirror. The top row shows filters with a gradual cut-off, the bottom row shows edge filters with sharp cut-offs.

Wide-field microscopy does not allow selection of a certain plane within the sample: the parallel light used for excitation in wide-field fluorescence microscopy penetrates through all layers of the sample. All molecules in all planes of the sample are excited, and their emission is focused by the objective and imaged onto the detector. The image on the detector is therefore a superposition of images of all individual planes in the sample: the position of the molecule in the z-direction that causes one particular spot is not encoded in the image.

25.3 CONFOCAL SCANNING MICROSCOPY

By *confocal microscopy*, optical selection of a specific layer of the sample becomes possible, and images of individual planes of the sample can be generated. In confocal microscopy, a parallel bundle of light is focused into the sample by an objective (Figure 25.8). The diffraction-limited focus in three dimensions is an ellipsoid, with its long axis along the beam in the z-direction, and a shorter axis in the xy-plane. Within this ellipsoid, the light intensity $I(x,y,z)$ decays exponentially in the x- and y-directions and in the z-direction from its center according to eq. 25.17:

$$I(x,y,z) = I_0 \exp\left(-\frac{2(x^2+y^2)}{\omega_1^2} - \frac{2z^2}{\omega_2^2}\right) \qquad \text{eq. 25.17}$$

I_0 is the light intensity in the center, ω_1 is the distance from the center where the intensity has decayed to I_0/e in the x- and y-directions, and ω_2 is the corresponding distance from the center in the z-direction. The volume defined by the surface where the intensity has decayed to I_0/e is called the *confocal volume*. Molecules that reside in the confocal volume experience high light intensity, are excited, and emit fluorescence. Fluorescence from the confocal volume is collected by the objective, focused into a small centric aperture, a *pinhole* (Figure 25.8), and eventually detected by a point detector such as an avalanche photodiode. The pinhole is the central element for the selection of a defined plane within the sample. Light emitted from the plane of interest is focused into the plane of the pinhole, and can therefore pass the pinhole completely. Light emitted from above or below is focused into different planes, is rejected by the pinhole, and does not reach the detector (Figure 25.8).

Confocal microscopy in its simplest form is used to measure fluorescence emission from a small volume within a sample. In fluorescence correlation spectroscopy, this mode is used to follow fluctuations in intensity over time to extract information on diffusion times (Section 25.6). To generate an image of the entire plane, we have to repeat the detection process at a different spot of the sample, piecing together the fluorescence intensity over the area of interest in the sample point by point and line by line. This technique is called *scanning confocal microscopy* (SCM). Either the confocal volume is scanned across the sample (beam scanning) or the sample is scanned relative to the fixed objective and confocal volume (sample scanning).

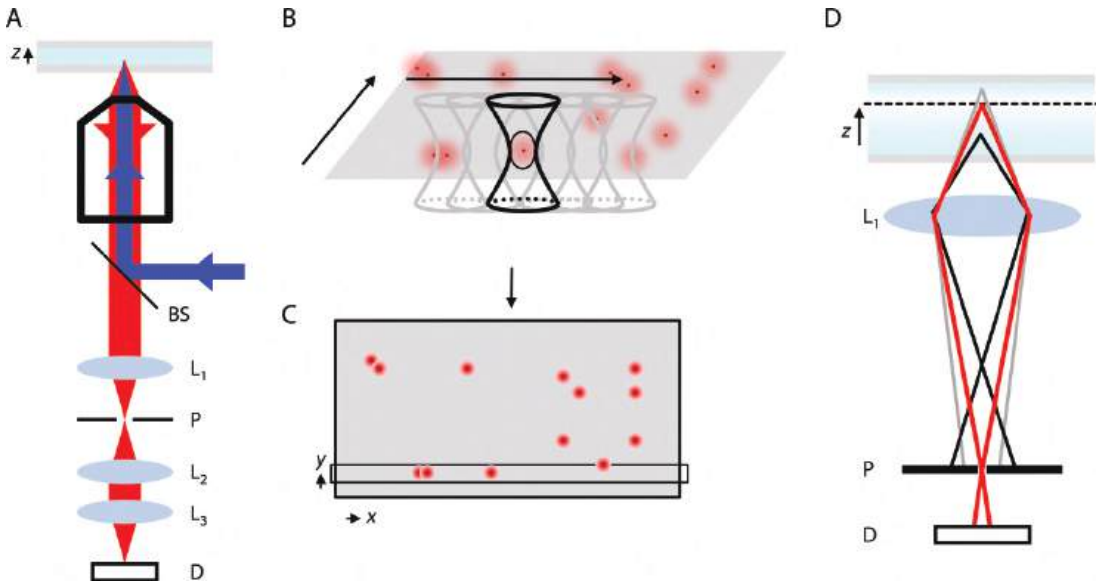


FIGURE 25.8: (Scanning) confocal microscopy. A: In a confocal microscope, parallel excitation light (blue) is coupled into a high numerical aperture objective and focused into a small volume, the *confocal volume*. Fluorophores within the volume are excited, and their fluorescence (red) is collected by the objective. The fluorescence is focused by lens L_1. A pinhole (P, a small circular aperture) is centered in the focal plane of L_1 and of the collecting lens L_2. The pinhole is the central element for optical selection of a specific plane within the sample. L_3 focuses the fluorescence light onto the point detector (D). B: To construct the image, fluorescence from each point has to be measured, one after the other. To achieve this, the focus (double cone for the beam waist, the confocal volume is indicated by the ellipse in the center) is scanned over the sample point by point and line by line. C: The scanned image shows an Airy disc for each emitter within the sample. D: optical selection for a plane within the sample. Light emitted from the plane of interest (black broken line; red) is focused into the aperture of the pinhole P by L_1. Light from planes above (gray) or below (black) is focused into different planes, and will be rejected by the pinhole. Only light from the plane of interest reaches the detector D.

Scanning confocal microscopy can also be used to generate three-dimensional images of specimen. To achieve this, imaging is performed layer by layer, and the optically selected focal plane is changed incrementally between two subsequent images. The three-dimensional image is then reconstructed from the set of two-dimensional images from different focal planes. The required scanning of the focus across the sample makes scanning confocal microscopy a rather slow imaging technique, limiting its applications. This limitation is overcome by *spinning disc confocal microscopy*. In this multiplexing technique, imaging is performed with a large number of pinholes instead of a single one, generating a set of foci at multiple points (Figure 25.9). The pinholes are arranged on a disc in such a way that rapid rotation of the disc during image acquisition enables the simultaneous excitation and fluorescence detection of entire rows of the image and decreases the acquisition time by typically two to three orders of magnitude.

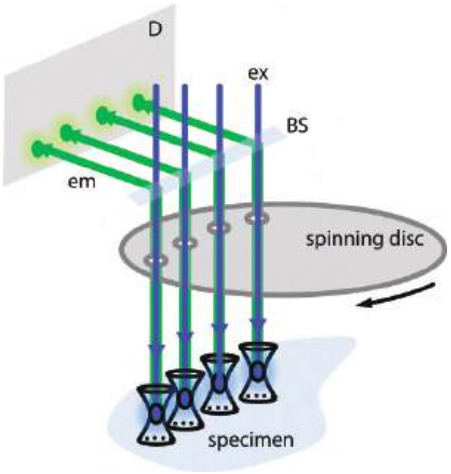


FIGURE 25.9: Spinning disc confocal microscopy. Excitation light (ex, blue) passes through a beam splitter (BS) onto a spinning disc containing an array of pinholes, generating a series of foci (blue) within the specimen. Emitted light (em, green) from the specimen is reflected by the beam splitter, and imaged onto a detector (D), giving rise to several image points simultaneously.

25.4 TOTAL INTERNAL REFLECTION MICROSCOPY

Wide-field microscopy allows simultaneous imaging of the entire surface area, but it does not select a specific plane in the *z*-direction. Using confocal microscopy, we can selectively image a layer of the sample, but we need to generate images point by point, which is time-consuming. *Total internal reflection fluorescence* (TIRF) *microscopy* combines the advantage of simultaneously detecting fluorescence from a larger area with selection of a certain layer of the sample (Figure 25.10). In TIRF microscopy, the incident laser beam is coupled into the objective off the optical axis, such that it will impinge at a defined angle on the interface between the slide surface and the sample (Figure 25.10). The refractive index n_2 of the sample (aqueous solution) is smaller than the refractive index n_1 of the glass or quartz slide. The excitation light therefore will be refracted away from the optical axis according to *Snellius' law* of refraction:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \qquad \text{eq. 25.18}$$

n_1 is the refractive index of the microscope slide, n_2 the refractive index of the sample, θ_1 and θ_2 are the incident and exiting angles of the beam with the optical axis. Above a certain angle θ, the *critical angle* θ_c, total internal reflection will occur. From Snellius' law, we can calculate the critical angle θ_c as

$$\theta_c = \text{arc}\sin\left(\frac{n_2}{n_1}\right) \qquad \text{eq. 25.19}$$

The beam will then not penetrate into the sample, but instead is reflected at the glass-water interface (Figure 25.10). Under these conditions, an *evanescent field* is generated on the distal side of the interface, within the sample. The intensity $I(z)$ of this evanescent field decays exponentially with increasing distance from the interface (see SPR in Section 29.3.1)

$$I(z) = I(0) \cdot e^{-z/d} \qquad \text{eq. 25.20}$$

where $I(0)$ is the intensity at the interface, z is the distance from the interface, and d is the decay parameter that depends on the wavelength λ of the excitation light, the refractive indices n_1 and n_2

of the slide and the sample, and on the angle θ at which the excitation light impinges on the glass-water interface:

$$d = \frac{\lambda_0}{4\pi} \cdot \frac{1}{\sqrt{n_1^2 \sin^2 \theta - n_2^2}} \qquad \text{eq. 25.21}$$

Fluorophores within a few 100 nm from the interface are within the range of the evanescent field and are excited, whereas fluorophores further away from the interface are not excited. TIRF microscopy therefore allows selective excitation of molecules on cell surfaces, or of molecules immobilized on the slide surface.

Instead of using an objective (objective-type TIRF microscopy), the conditions for total internal reflection can also be generated using prisms (prism-type TIRF microscopy; Figure 25.10). In both cases, the fluorescence emission is collected by an objective and imaged onto a CCD camera.

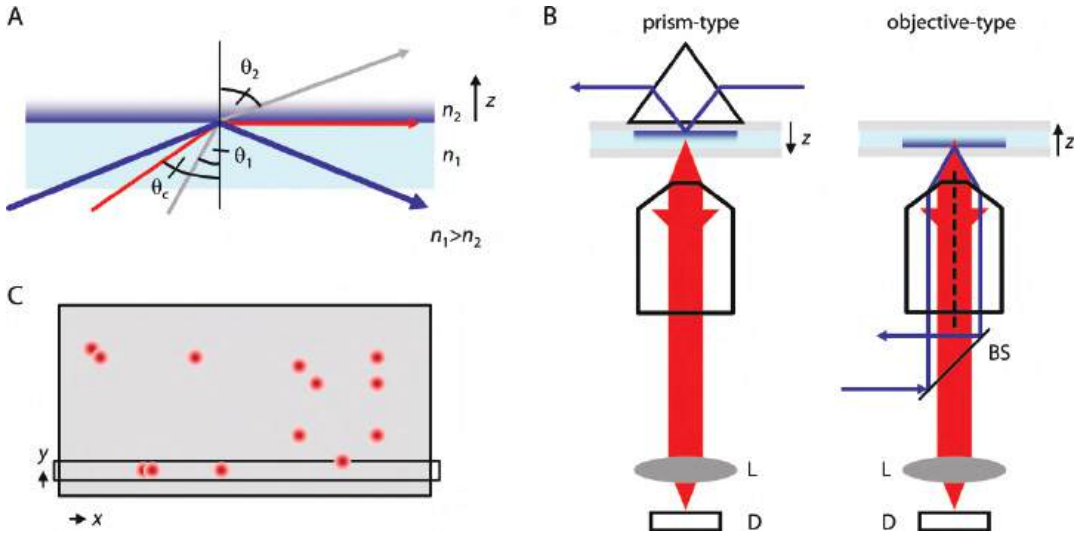


FIGURE 25.10: TIRF microscopy. A: Total internal reflection. When light passes the interface from an optically dense medium (glass/quartz) with a refractive index n_1 to a medium with lower refractive index ($n_2 < n_1$; water), it is refracted away from the optical axis ($\theta_2 > \theta_1$) according to Snellius' law (eq. 25.18; red). At incident angles above a critical angle θ_c, the light will not enter the second medium, but is reflected at the interface (blue). At the critical angle θ_c, the light beam travels along the interface (gray). Total internal reflection leads to an evanescent field on the distal side of the interface (blue gradient), and to light intensity close to the surface that decays exponentially along the *z*-axis. As a consequence, fluorophores at the surface will be excited and emit light. B: Prism-type and objective-type TIRF microscopes. In the prism-type TIRF microscope, the critical angle is achieved by passing the excitation light through a prism. The evanescent field is generated at the top interface (blue gradient). In objective-type TIRF microscopy, it is achieved by coupling the excitation light into the objective off the optical axis. In this case, the evanescent field is generated at the bottom interface (blue gradient). Fluorescence from molecules at the surface is collected by an objective and imaged onto an area detector D (BS: beam splitter, L: lens). C: The image shows an Airy disc for each emitter on the surface.

Sectioning of the sample can also be achieved by *light-sheet microscopy*, where a section of the sample, perpendicular to the direction of observation or inclined with respect to the optical axis, is selectively illuminated (Box 25.2). Light-sheet microscopy is not limited to selective excitation near surfaces and avoids artifacts from surface effects.

BOX 25.2: OPTICAL SECTIONING OF SAMPLES BY LIGHT-SHEET ILLUMINATION.

In light-sheet microscopy, the excitation light is passed through a cylindrical lens. This lens only focuses the light in one direction (see Section 25.1.1), leading to the generation of a thin light sheet that illuminates a thin area within the sample (Figure 25.11). In *highly inclined and laminated optical sheet microscopy* (HILO; Figure 25.11), a variation of TIRF microscopy, the excitation light is coupled into the objective off the optical axis, but such that it impinges on the interface between the glass slide and the sample at an angle below the critical angle for total internal reflection. The light is refracted away from the optical axis at the glass-water interface and passes through the sample at an acute angle with respect to the surface. Light-sheet microscopy does not reach the signal-to-noise ratio of TIRF microscopy, but offers the advantage of selectively observing molecules in a narrow plane away from the microscope slide surface. The illumination of a limited area within the sample reduces photodamage. Tokunaga *et al.* (2008) Nat. Methods 5(2): 159–161.

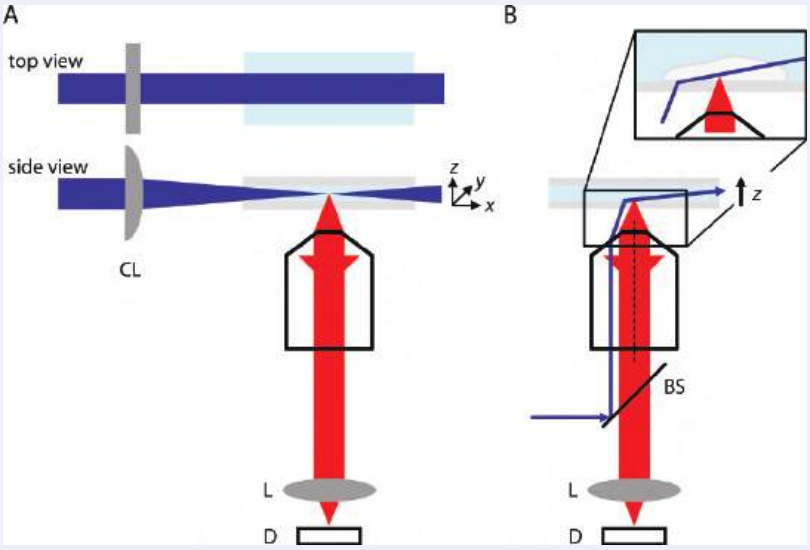


FIGURE 25.11: Light-sheet microscopy. A: Light is focused in one dimension only by a cylindrical lens (CL), leading to illumination of the sample by a sheet of light in the *xy*-plane. Fluorescent molecules within this plane are excited and their fluorescence is imaged onto a detector (D) by a lens (L). B: In highly inclined and laminated optical sheet microscopy (HILO), the light sheet is generated at an acute angle with the surface by illuminating the glass-water interface at an angle larger than the critical angle for total internal reflection. Only an inclined plane within the specimen (inset, gray) is observed. BS: beam splitter, L: lens, D: detector.

Fluorescence microscopes can be equipped with multiple excitation lasers to excite different fluorophores and to image two or more differentially labeled molecules simultaneously. Images obtained from the same specimen using different fluorophores can then be superimposed. Appearance of fluorescence from different molecules in the same position within a cell is often used as evidence for colocalization and a possible functional interaction. In *FRET microscopy* (see Section 21.5.10), the donor fluorophore is excited, and the emission from donor and acceptor fluorophore is separated using a dichroic mirror. Focusing the donor and acceptor fluorescence onto separate detectors, either two avalanche photodiodes for confocal microscopy or next to each other on the CCD chip of the camera in wide-field and TIRF microscopy, provides donor and acceptor fluorescence as a function of time (confocal microscopy) or space (wide-field and TIRF microscopy). The FRET efficiency can be calculated for each time point, or pixel by pixel as a function of position, providing information on the

conformational fluctuations of a molecules carrying two dyes or on colocalization and interaction of two fluorescently labeled entities. We will come back to FRET microscopy in Section 25.7.6.

25.5 FLUORESCENCE LIFETIME IMAGING MICROSCOPY

Wide-field, scanning confocal, and TIRF microscopy yield images of the sample, *i.e.* fluorescence intensity as a function of position. Sometimes it is important to detect and distinguish two different fluorophores or to detect different environments of the same fluorophore, which is not necessarily possible from fluorescence intensities. If a pulsed excitation source is used, the fluorescence lifetime at each spot of the sample can be measured (see Section 21.5.9). Different fluorophores can then be distinguished by their lifetimes. *Fluorescence lifetime imaging microscopy* (FLIM) requires spatial resolution to generate the image, combined with temporal resolution on the nanosecond time scale at each spot to measure the fluorescence decay. Such measurements can be achieved by using a *streak camera* as the detector. Just like any conventional photodetector, a streak camera converts incoming photons into secondary (photo)electrons. Streak cameras can imprint information on the arrival time of the photon on the photoelectron. This is achieved by passing the photoelectrons through an electric field whose intensity varies with time (Figure 25.12). Photoelectrons are deflected when they pass this electric field, and the extent of the deflection depends on the magnitude of the electric field at this time point. If the magnitude of the electric field increases with time, electrons that arrive earlier will then be deflected less than electrons that pass the electric field later. In other words, electrons are deflected proportionally to their arrival time. The time-dependent deflection in one direction thus introduces a second axis that reflects the arrival time of the photon.

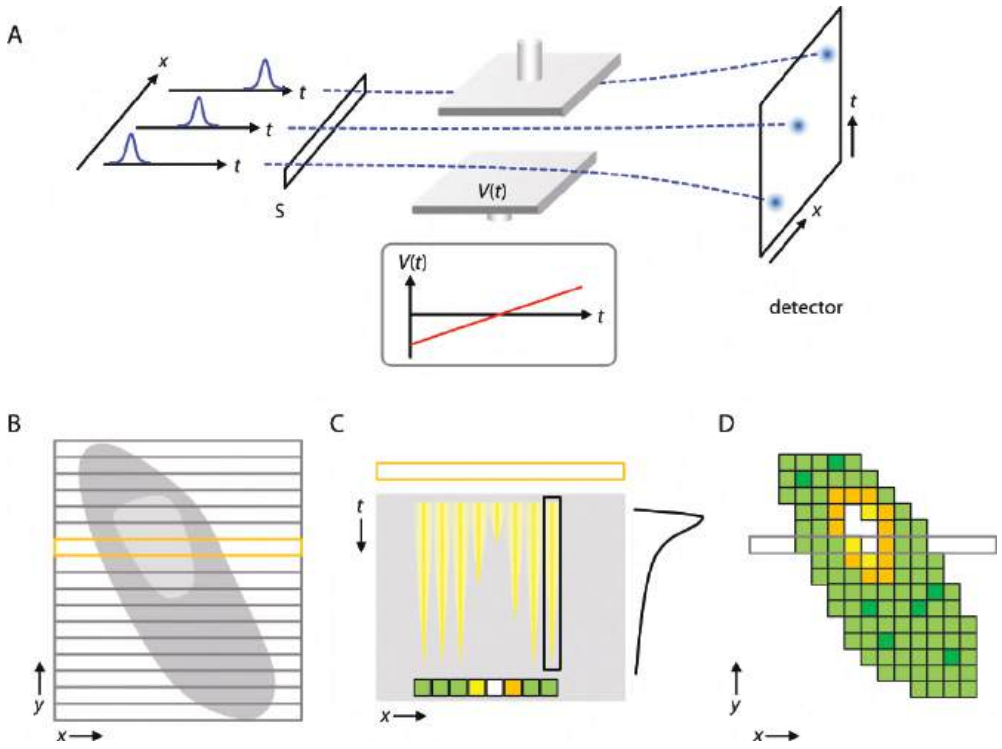


FIGURE 25.12: Streak camera and fluorescence lifetime imaging. A: Principle of a streak camera. Photons arriving at different positions along the *x*-axis and at different time points *t* pass the entry slit of the streak camera and generate photoelectrons that are accelerated toward the detector. Along the way, they pass a set of deflection plates. A time-dependent voltage leads to a time-dependent deflection of the electrons. The electrons arrive at the detector at their *x*-position and are spread along a second axis according to their arrival times *t*. B: A cell (nucleus in light gray, cytoplasm in dark gray) is imaged line by line. C: Imaging of the highlighted line gives a streak image with photon arrival times along the *t*-axis and their position along the *x*-axis. From the time-dependent photon count, fluorescence decays and fluorescence lifetimes are calculated at each spot in the line. The lifetime is color-coded (green-orange-yellow-white from large to small lifetimes). D: The lifetime image is constructed from the lifetimes measured at each spot per line for all lines.

The electrons then pass a microchannel plate (Section 21.5.9; Figure 29.11 in Section 29.1.3) for amplification of the signal, and are detected by a phosphor screen. The light from the phosphor screen is imaged onto a CCD camera. One round of measurements thus yields one line of the image. Line by line, the complete image is then built up, yielding a three-dimensional dataset with x- and y-axes and an additional t-axis with the fluorescence decay at each x,y-coordinate. The fluorescence lifetime can be calculated from the fluorescence decay along the t-axis. The whole dataset is represented as a color-coded image in which the color corresponds to the lifetime at a particular spot (Figure 25.12).

Since fluorescence lifetimes are sensitive to pH, to the local oxygen concentration, or to the presence of fluorescence acceptors, FLIM can provide spatially resolved information on these parameters within cells.

25.6 FLUORESCENCE (CROSS-)CORRELATION SPECTROSCOPY

We have seen in Section 25.3 that confocal microscopy can be used to construct fluorescence images by scanning the confocal volume across the sample. Instead of measuring fluorescence intensities as a function of position, we can also measure fluorescence from the confocal volume as a function of time. The measured fluorescence will fluctuate around its mean value because of diffusion (see Section 4.1), and from the fluctuation speed, we can gain information on diffusion velocities.

25.6.1 Fluorescence correlation spectroscopy

The confocal volume is tiny (on the order of femtoliter), and at any given time we will have a limited number of fluorescent molecules within the focus. At sufficiently low concentrations, the average number of fluorescent molecules is on the order of ten. Because of Brownian motion (see Section 4.1), fluorescent molecules constantly enter and leave the confocal volume, leading to fluctuations in the average number of molecules within the confocal volume, and a fluctuation in the fluorescence intensity. At moderate concentrations of the fluorescent molecule, the fluorescence will therefore measurably fluctuate over time around the mean intensity. The fluctuation $\delta I(t)$ is the difference between the fluorescence intensity $I(t)$ and the mean fluorescence intensity $\langle I \rangle$ (averaged over time):

$$\delta I(t) = I(t) - \langle I \rangle \qquad \text{eq. 25.22}$$

The *autocorrelation function* $G(\tau)$ (see Section 23.1.2) relates the fluorescence intensity at time t, $I(t)$, to the fluorescence at a later time $t+\tau$, $I(t+\tau)$:

$$G(\tau) = \frac{\langle I(t) \cdot I(t+\tau) \rangle}{\langle I \rangle^2} \qquad \text{eq. 25.23}$$

The angle brackets indicate that we calculate the mean of the respective parameters over time. We express the intensity at each point in time as the mean intensity $\langle I(t) \rangle$ plus the fluctuation $\delta I(t)$, and obtain

$$G(\tau) = 1 + \left(\frac{\langle (\langle I \rangle + \delta I(t)) \cdot (\langle I \rangle + \delta I(t+\tau)) \rangle}{\langle I \rangle^2} \right) \qquad \text{eq. 25.24}$$

We can factorize the numerator to

$$G(\tau) = \left(\frac{\langle \langle I \rangle \langle I \rangle + \langle I \rangle \cdot \langle \delta I(t) \rangle + \langle I \rangle \cdot \langle \delta I(t+\tau) \rangle + \langle \delta I(t) \cdot \delta I(t+\tau) \rangle \rangle}{\langle I \rangle^2} \right) \qquad \text{eq. 25.25}$$

The means of the fluctuations, $\langle \delta I(t) \rangle$ and $\langle \delta I(t+\tau) \rangle$, over time are zero, and their products with $\langle I \rangle$ are also zero. The autocorrelation function $G(\tau)$ thus becomes

$$G(\tau) = 1 + \left(\frac{\langle \delta I(t) \cdot \delta I(t+\tau) \rangle}{\langle I \rangle^2} \right) \qquad \text{eq. 25.26}$$

The autocorrelation curve decays from its initial value to 1. It is often plotted with a logarithmic τ-axis, resulting in a sigmoidal shape (Figure 25.13). The inflection point of the curve marks the average diffusion time τ_D of the molecule through the confocal volume. The amplitude of the autocorrelation curve is inversely proportional to the average number N of molecules in the confocal volume.

The second term in eq. 25.26 summarizes the time-dependent contributions to the autocorrelation function. This time-dependent part decays from the starting value to zero. The magnitude of the autocorrelation at time τ reflects the decreasing probability of detecting a photon at increasing time τ after detection of a photon at $\tau = 0$.

The time the molecule spends within the confocal volume depends on its translational diffusion coefficient D_t (Section 4.1) and on the dimensions of the confocal volume. We have seen before that the confocal volume is an ellipsoid with the short half axis ω_1 in the x- and y-directions, and long half axis ω_2 along the z-direction (Figure 25.13), and the light intensity as a function of x, y, and z is

$$I(x,y,z) = I_0 \exp\left(-\frac{2(x^2+y^2)}{\omega_1^2} - \frac{2z^2}{\omega_2^2} \right) \qquad \text{eq. 25.17}$$

Taking into account diffusion as the (only) source of the measured fluctuations, the autocorrelation curve can be described by

$$G(\tau) = 1 + \frac{1}{N} \left(\frac{1}{1 + 4D_t\tau/\omega_1^2} \right) \left(\frac{1}{1 + 4D_t\tau/\omega_2^2} \right)^{0.5} \qquad \text{eq. 25.27}$$

N is the average number of molecules in the confocal volume, D_t is the diffusion coefficient, and ω_1 and ω_2 are the distances from the optical axes in the x/y- and z-directions at which the light intensity has decayed by $1/e$ from the maximum. Note that the τ remaining in eq. 25.27 is the autocorrelation time, not the diffusion time! Thus, from autocorrelation curves we obtain diffusion times (inflection point) and the number of molecules within the confocal volume (inverse of the amplitude). We can convert these values to diffusion coefficients and concentrations only if we know the dimensions of the confocal volume. To determine diffusion coefficients and concentrations from autocorrelation curves, we describe the experimentally obtained curves by eq. 25.27, with D_t and N as fitting parameters and ω_1 and ω_2 as fixed parameters that define the dimensions of the confocal volume.

Diffusion of molecules into and out of the confocal volume is not the only process that leads to fluctuations of the fluorescence intensity, though. Fluctuations can also be caused by the photophysical properties of the fluorophore, such as transitions to the triplet state, or by chemical reactions and binding events. If polarized light is used for excitation, rotational motion of the fluorophores also causes fluorescence fluctuations. Any process that causes fluctuations of the fluorescence intensity can be studied by *fluorescence correlation spectroscopy* (FCS) (Figure 25.13).

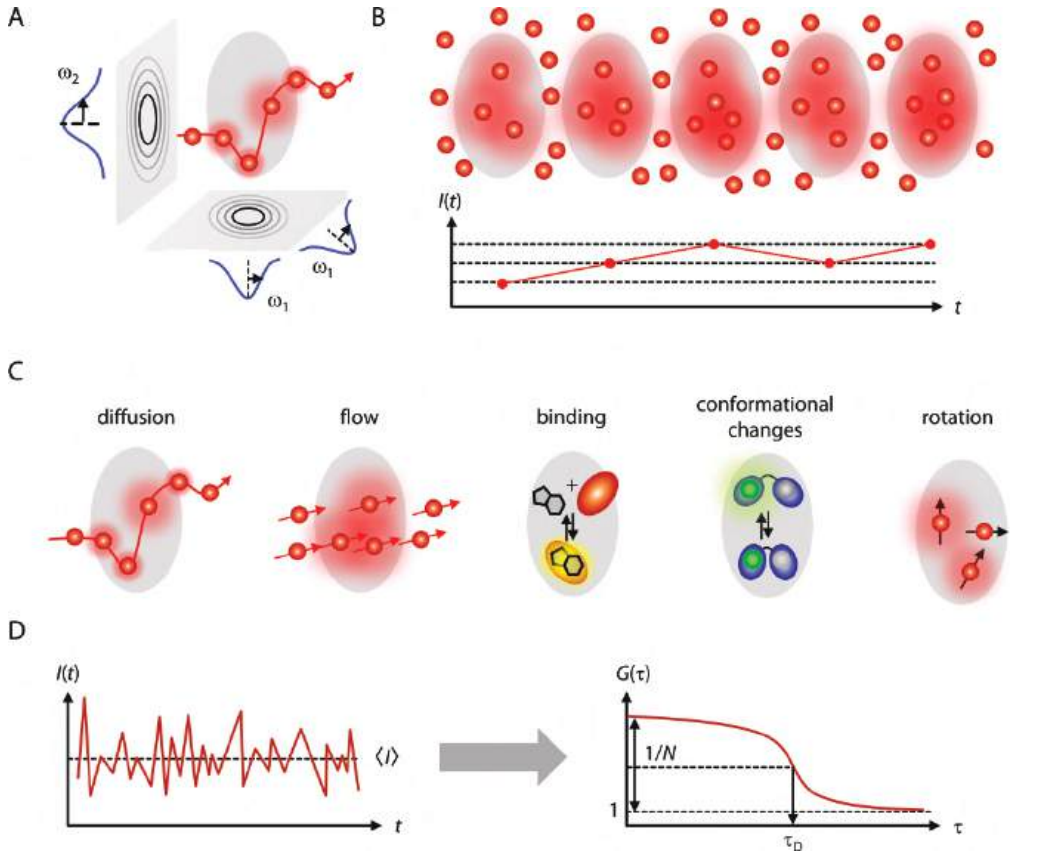


FIGURE 25.13: Fluorescence correlation spectroscopy (FCS). A: Dimensions of the confocal volume. The ellipsoidal volume can be described in terms of the half-axes ω_1 (in the *x*- and *y*-directions) and ω_2 (in the *z*-direction). B: Principle of FCS: diffusion of molecules into and out of the confocal volume leads to fluctuations of the measured fluorescence. C: Processes that lead to fluctuation: diffusion and flow alter the number of molecules within the confocal volume and cause thereby fluorescence fluctuations. Binding events or conformational changes can lead to changes in fluorophore brightness and cause fluorescence fluctuations. If polarized light is used for excitation, the rotation of molecules within the confocal volume also leads to fluorescence fluctuations. D: Measurement (left) and autocorrelation curve (right). The amplitude of the sigmoidal autocorrelation curve $G(\tau)$ is proportional to the inverse average number of molecules ($1/N$) within the confocal volume. The inflection point corresponds to the diffusion time τ_D.

In principle, FCS measurements require only one detector that measures fluorescence from the confocal volume as a function of time. However, any noise from this detector will also correlate with itself and contribute to the autocorrelation curve. To avoid contributions from detector noise, FCS measurements are often performed by splitting the fluorescence signal and focusing each half onto a separate detector. The correlation curve obtained from the two halves of the signal then only contains information from correlated fluctuations, whereas the uncorrelated noise from the two detectors cancels.

25.6.2 FCS TO MONITOR BINDING EVENTS

FCS can be used to monitor (high affinity) ligand binding by changes in diffusion time. Typically, increasing concentrations of the binding partner with the higher molecular mass will be added to a fluorescently labeled smaller ligand. The short diffusion time of the free ligand increases when it binds to the macromolecule. This increase in diffusion time serves as a probe for binding. In FCS, typical ligand concentrations are in the low nanomolar range, and K_d values in the high picomolar to low nanomolar range can be determined reliably. There are only few other techniques that can quantify such high-affinity interactions.

When we measure binding by FCS, two processes occur during our measurement: diffusion leads to fluorescence fluctuations, and binding and dissociation alter the diffusion coefficient. Binding and dissociation can be fast or slow on the time scale of the diffusion. If binding and dissociation are slower than diffusion ($\tau_{binding} \gg \tau_D$), no binding or dissociation occurs during the passage of the molecules through the confocal volume: molecules will diffuse in and out in their free form, with the diffusion coefficient D_{free}, or they enter and exit in the bound state and diffuse more slowly with D_{bound}. The autocorrelation curve therefore shows two inflection points, one corresponds to the diffusion of the free ligand, the second corresponds to the slower diffusion of the bound ligand (Figure 25.14). The autocorrelation can be described in terms of the two diffusion coefficients D_t^{free} and D_t^{bound} and the fractions of free and bound ligand, α and 1–α, respectively:

$$G(\tau) = 1 + \frac{1}{N}\left(\alpha X_{free} + (1-\alpha) X_{bound}\right) \qquad \text{eq. 25.28}$$

with

$$X_{free} = \left(\frac{1}{1+4D_t^{free}\tau/\omega_1^2}\right)\left(\frac{1}{1+4D_t^{free}\tau/\omega_2^2}\right)^{0.5} \qquad \text{eq. 25.29}$$

and

$$X_{bound} = \left(\frac{1}{1+4D_t^{bound}\tau/\omega_1^2}\right)\left(\frac{1}{1+4D_t^{bound}\tau/\omega_2^2}\right)^{0.5} \qquad \text{eq. 25.30}$$

The diffusion coefficients for the free and bound ligand can be determined in reference measurements. The fractions bound and free for each curve can then be determined from the amplitudes of the corresponding component of the autocorrelation curve (Figure 25.14). Plotting the fraction bound as a function of the concentration of the added macromolecule yields a binding curve that can then be analyzed for the K_d value (eq. 21.69 in Section 21.5.5.1).

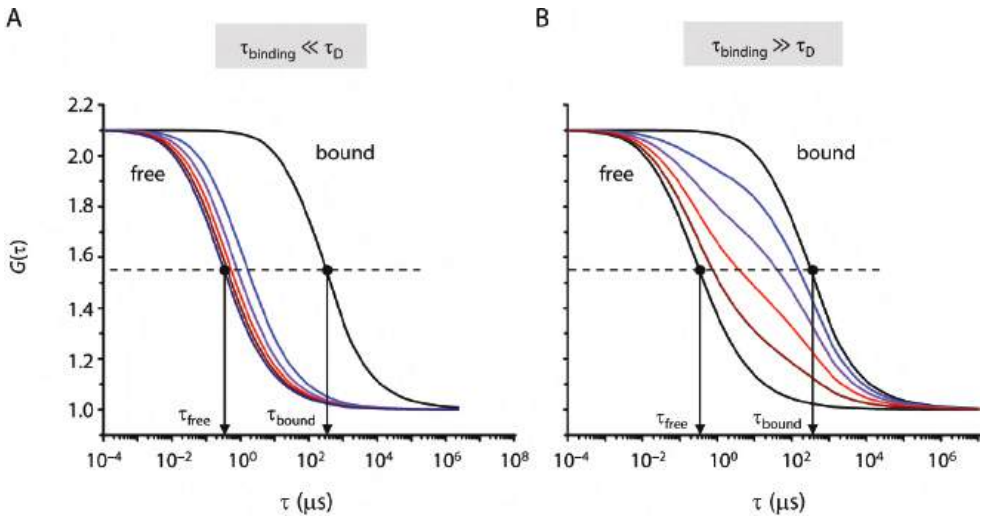


FIGURE 25.14: FCS to monitor binding and to determine K_d values. A: Autocorrelation curves for a rapid equilibrium ($\tau_{binding} \ll \tau_D$) from 0% binding (free ligand, black) over 20% (dark red), 40% (red), 60% (violet), 80% (blue) to 100% bound ligand (black). From the autocorrelation curves, a mean diffusion time τ_D can be determined. This mean diffusion time represents the population-weighted average of the diffusion times τ_{free} and τ_{bound} for free and bound ligand. From the mean diffusion time, the fraction free and bound ligand can be determined. B: Autocorrelation curves for a slow equilibrium ($\tau_{binding} \gg \tau_D$). The autocorrelation curves are population-weighted linear combinations of the curves for free and bound ligand. From the amplitudes of the two components of the measured curve, the fraction free and bound ligand can be determined.

If the binding reaction is fast ($\tau_{binding} \ll \tau_D$), binding and dissociation are in dynamic equilibrium while the molecules reside in the confocal volume. In this case, the set of autocorrelation curves of a titration experiment show a single inflection point that continuously shifts toward longer times for increasing concentrations of macromolecule (Figure 25.14). This *apparent diffusion coefficient* $\langle D_t \rangle$ from each autocorrelation curve is the population-weighted average of the diffusion times for the free and bound ligand.

$$G(\tau) = 1 + \frac{1}{N}\left(\frac{1}{1+4\langle D_t\rangle\tau/\omega_1^2}\right)\left(\frac{1}{1+4\langle D_t\rangle\tau/\omega_2^2}\right)^{0.5} \qquad \text{eq. 25.31}$$

To obtain a binding curve, the apparent diffusion coefficient is plotted as a function of the concentration of the macromolecule, and the K_d value can be determined by data fitting (see eq. 21.69).

For the quantitative analysis of FCS experiments to determine diffusion coefficients or numbers of molecules, the size and shape of the confocal volume have to be known, which is often difficult to achieve and experimentally difficult to control. These problems for quantitative FCS can be circumvented by using *dual-focus FCS* (2f-FCS). Here, two overlapping foci are generated that are laterally shifted by a known distance. The distance between the centers of the foci provides an external length ruler for the exact determination of diffusion coefficients (Box 25.3).

BOX 25.3: DUAL-FOCUS FCS (2F-FCS) AND ABSOLUTE DIFFUSION COEFFICIENTS.

2f-FCS is based on two confocal volumes that overlap but are laterally shifted by a known distance. To generate the two foci, two excitation lasers with orthogonal polarization are used. Their beams are passed through a Nomarski prism that generates a lateral shift between the two components because of their difference in polarization. An objective then focuses the two beams in separate spots (Figure 25.15). The distance between the two foci can be determined in reference measurements with compounds of known diffusion coefficients, or by imaging the two foci with fluorescent beads. The inter-focal distance depends only on the type of prism and on the objective used. From the autocorrelation curves of fluorescence from focus 1 or from focus 2, as well as the cross-correlation between fluorescence from both foci (see Section 25.6.3, Figure 25.15), the diffusion coefficient can be determined by global analysis. 2f-FCS allows the determination of diffusion coefficients *in vivo*. 2f-FCS has been used to measure lipid diffusion in black lipid membranes. In combination with line-scanning, ligand affinities of membrane receptors have been determined within the membrane of living cells. Weiss & Enderlein (2012). Chem. Phys. Chem. 13(4): 990–1000. Dorlich *et al.* (2015) Sci. Rep. 5: 10149.

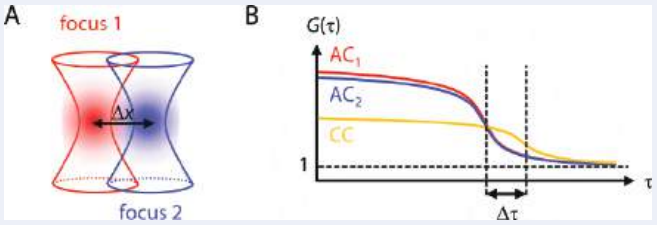


FIGURE 25.15: 2f-FCS. A: Principle. Two overlapping, laterally shifted foci serve as external ruler. B: Autocorrelation curves from each focus (AC_1, AC_2) and cross-correlation curve (CC) from both foci. The cross-correlation curve is shifted to longer times τ because of the distance between the foci. Global analysis of auto- and cross-correlation curves permits determination of absolute diffusion coefficients.

The size of the confocal volume depends on the wavelength of the light used, the width of the incoming beam, and the numerical aperture of the objective. In confocal scanning microscopy, higher spatial resolution is achieved with smaller foci. Although the size of the confocal volumes can be reduced somewhat by widening the incoming excitation beam or by increasing the numerical aperture of the objective, a much larger reduction in size is achieved by using *two-photon excitation* (Box 25.4).

BOX 25.4: FCS WITH TWO-PHOTON EXCITATION.

Instead of excitation of a fluorophore by illumination with light corresponding to the energy difference between S_0 and S_1 states, excitation can in principle also be achieved by absorption of two photons of twice the wavelength that each provide half of the energy for the transition (Figure 25.16). The simultaneous absorption of two photons is an unlikely process, and high light intensities are required to achieve two-photon excitation. As a consequence, two-photon excitation can only occur close to the center of the confocal volume. The effective confocal volume for two-photon excitation is therefore much smaller. A smaller confocal volume provides increased resolution for confocal imaging. Excitation with infrared light instead of visible light is advantageous for imaging of cells: the background fluorescence is lower, and photodamage to the cell is reduced.

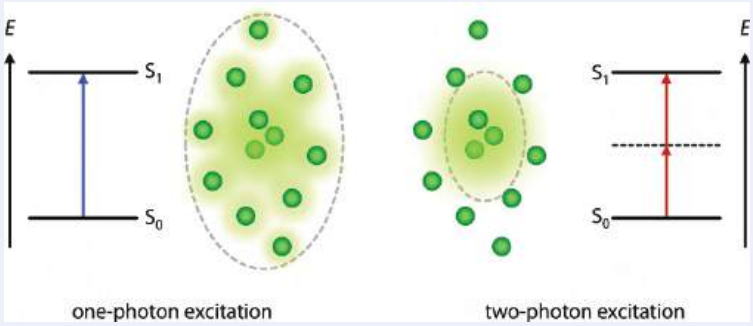


FIGURE 25.16: One- and two-photon excitation. Left: A green fluorophore can be excited by illumination with blue light whose energy corresponds to the energy difference between the S_0 and S_1 states. The excitation is most efficient in the center of the confocal volume and decays exponentially in the *x*-, *y*-, and *z*-directions. Right: Alternatively, a green fluorophore can be excited by absorption of two red photons of half the energy. The dotted line indicates a virtual intermediate state between S_0 and S_1. The absorption of two photons is unlikely and requires higher light intensities than excitation by one photon. As a consequence, the fluorophore is only excited efficiently close to the center of the confocal volume, but not at its rim. The effective confocal volume (broken gray line) is significantly reduced compared to one-photon excitation.

While information on rates of diffusion, photophysical processes, and chemical reactions is determined from autocorrelation curves, information on the number of molecules and their brightness is typically obtained from *photon counting histograms* (PCH) that count how often a certain intensity has been measured (Figure 25.17). The PCH reveals the presence of different species with different brightness and their relative populations, and thereby allows detection of dimer formation or oligomerization.

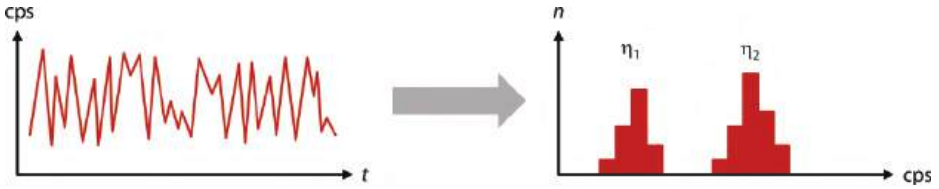


FIGURE 25.17: Photon counting histograms, number of molecules, and molecular brightness. The fluctuating fluorescence signal from the confocal volume (left) is replotted in terms of the frequency with which a certain count rate (cps, counts per second) has been detected (photon counting histogram, PCH; right). The PCH shows two species, one with a low brightness η_1 (left peak) and a second with a higher brightness η_2 (right peak). These species could be a monomer and a dimer. Note that photon counting statistics leads to more complex shapes and more complex analyses of photon counting histograms.

25.6.3 Fluorescence cross-correlation spectroscopy

Instead of correlating fluctuations of a fluorescence signal with itself, it is also possible to correlate the signal from two fluorophores with each other. In *fluorescence cross-correlation spectroscopy* (FCCS), two dyes are used. The fluorescence emitted from molecules in the confocal volume is split into the spectral regions of the two dyes for quantification on two separate detectors. The detected signals $I_1(t)$ and $I_2(t)$ are then cross-correlated to calculate the *cross-correlation function* $G(\tau)$:

$$G(\tau) = 1 + \frac{\langle \delta I_1(t) \cdot \delta I_2(t+\tau) \rangle}{\langle I_1 \rangle \cdot \langle I_2 \rangle} \qquad \text{eq. 25.32}$$

The shape of the cross-correlation function is identical to the autocorrelation function, but the meaning is somewhat different. The amplitude is now proportional to the number of species in the confocal volume that carry both fluorophores, the inflection point corresponds to the diffusion time of the species that carries both fluorophores. The cross-correlation curve describes the probability to detect a photon in channel 2 at time point τ after detection of a photon in channel 1 at $\tau = 0$ (Figure 25.18).

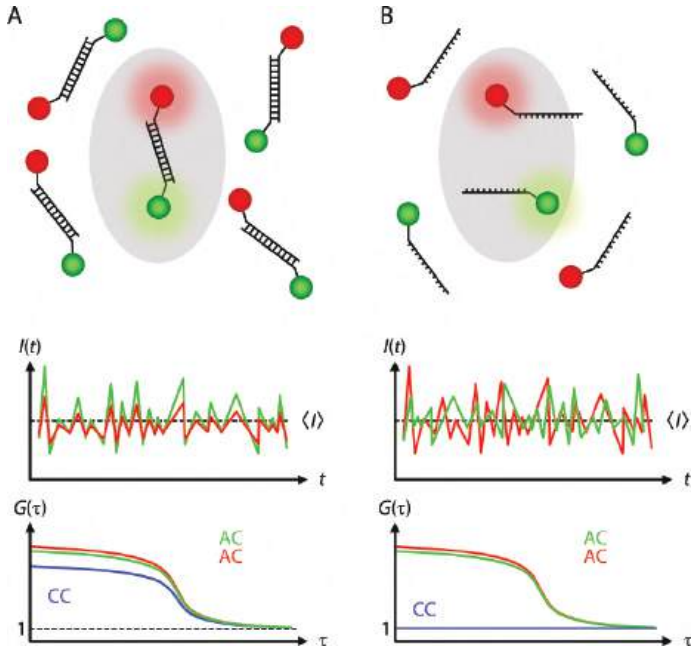


FIGURE 25.18: Fluorescence cross-correlation spectroscopy (FCCS). A: When two fluorophores diffuse together, their fluorescence fluctuations are correlated. Whenever green fluorescence is decreasing, red fluorescence also decreases, and *vice versa*. The cross-correlation therefore has a measurable amplitude. The inflection points of the cross-correlation function as well as of the autocorrelation functions for green and red fluorescence are identical and correspond to the diffusion time of the entity that carries both fluorophores. B: If the fluorophores belong to two non-interacting molecules, the green and red fluorescence fluctuations are independent of each other and uncorrelated. The amplitude of the cross-correlation function is zero. The diffusion time of each molecule can be extracted from the inflection point of the green and red autocorrelation functions.

FCCS can be applied to study interactions of molecules both *in vitro* and *in vivo* (Box 25.5). Because of the high sensitivity of confocal fluorescence microscopy, high-affinity interactions can be studied.

BOX 25.5: EXAMPLES OF FC(C)S STUDIES *IN VIVO*.

FCCS has been used to study the interaction between fluorescently labeled transcription factors jun and fos *in vivo*. Jun and fos were fused to the fluorescent proteins EGFP and mRFP, respectively, and overproduced in living cells. Cells producing an EGFP-mRFP fusion protein served as a positive control, cells producing EGFP and mRFP individually were used as a negative control. Cross-correlation curves with significant amplitudes were obtained from cells containing jun-EGFP and fos-mRFP, demonstrating their heterodimerization *in vivo*. When the dimerization domains of both proteins were deleted, the cross-correlation amplitude was lost, demonstrating the loss of dimerization. Baudendistel *et al.* (2005) Chem. Phys. Chem. 6(5): 984–990.

An *in vivo* FCS/FCCS study of the RNA-induced silencing complex (RISC), a protein-RNA complex involved in RNA interference, has shown that this complex exists in a cytoplasmic and a nuclear form with different diffusion coefficients. Nuclear RISC (ca. 160 kDa) diffuses much more rapidly than the ca. 3 MDa cytoplasmic RISC. Together with biochemical data, this study provided evidence for a mechanism in which the Argonaute protein Ago2 and a small regulatory RNA form a RISC complex in the cytoplasm that is then shuttled to the nucleus where it interacts with its target mRNA and mediates RNA-dependent regulation of gene expression. Ohrt *et al.* (2008) Nucleic Acids Res. 36(20): 6439–6449.

FCS (and FCCS) can also be used in high-throughput format to study the dynamics of large numbers of proteins in living cells. Using automated screening and time-lapse acquisition of FCS data at specific subcellular locations, and automated data analysis, the dynamics of more than fifty nuclear proteins has been studied by high-throughput FCS. Wachsmuth *et al.* (2015) Nat. Biotechnol. 33(4): 384–389.

25.7 SINGLE-MOLECULE FLUORESCENCE MICROSCOPY

25.7.1 Principles of single-molecule microscopy

In Sections 25.2–25.5 we have treated different types of fluorescence microscopy to generate images of tissues, cells, and sub-cellular structures. Fluorescence microscopy with laser excitation and ultrasensitive detectors provides the sensitivity to detect fluorescence from single molecules. Wide-field, confocal scanning, and TIRF microscopy can be used to obtain images of single molecules immobilized on surfaces (Box 25.6, see also Section 29.3.4). Each emitting molecule gives rise to a bright spot in the image of the surface. Integration over the entire spot yields the fluorescence intensity of the particular molecule. In contrast, FCS-type confocal microscopy allows measurement of fluorescence from single molecules in solution as a function of time while they are traversing the confocal volume. Each molecule gives rise to a burst of fluorescence photons that is detected by an avalanche photodiode.

BOX 25.6: SURFACE IMMOBILIZATION OF SINGLE MOLECULES FOR TIRF MICROSCOPY.

Microscope slides for surface immobilization of single molecules need to be passivated to minimize unspecific interactions of the molecule of interest with the surface. Passivation is achieved with the protein bovine serum albumin (BSA) that interacts non-specifically with

(*Continued*)

the glass surface, with polyethylene glycol (PEG) that forms polymer brushes on the glass slide, or by depositing a lipid bilayer on the glass surface. Surface immobilization for single-molecule fluorescence studies is often performed using the biotin-streptavidin interaction (Figure 25.19). A streptavidin functionalization of the surface is achieved by doping the passivation layer on the microscope surface with biotinylated BSA, PEG, or lipids to which streptavidin is coupled. Proteins are biotinylated either *in vivo* or *in vitro*, and immobilized on the streptavidin-functionalized surface. Biotinylated nucleic acids are either chemically synthesized (DNA, RNA), biotinylated by adding a biotinylated starter nucleotide to *in vitro* transcription reactions (RNA), or generated by ligation or hybridization of a biotinylated oligonucleotide (DNA, RNA). Alternatively, proteins and nucleic acids can be incorporated into lipid vesicles that are surface-immobilized. This method minimizes surface effects because the molecule of interest is in solution within the vesicle, yet spatially confined for long-time observation. Vesicles containing protein pores allow addition of small molecules to the molecule of interest, but the addition of larger interaction partners is not possible.

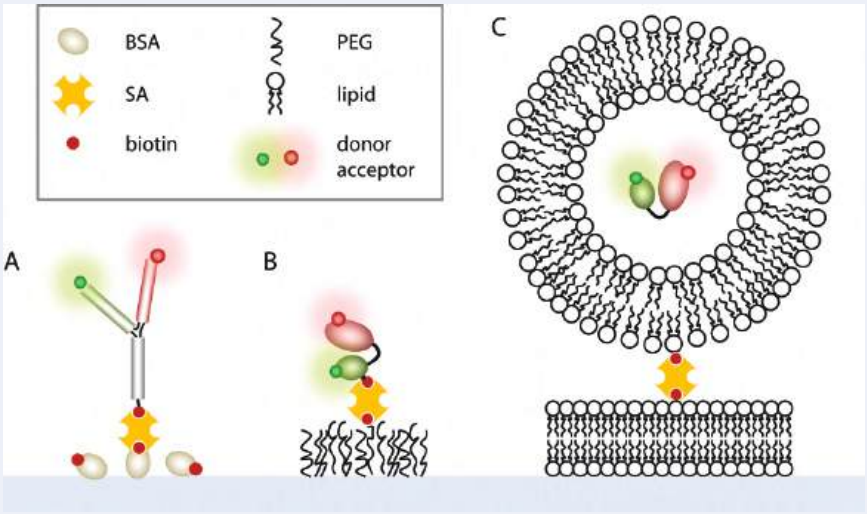


FIGURE 25.19: Immobilization schemes for proteins and nucleic acids. A: Immobilization of biotinylated nucleic acids (cylinders) on glass surfaces functionalized and passivated with biotinylated bovine serum albumin (BSA). B: Immobilization of biotinylated proteins (ovals) on polyethylene glycol-passivated glass surfaces doped with biotin. C: Incorporation of proteins (or nucleic acids) into lipid vesicles and immobilization of lipid vesicles on surfaces with biotin-doped lipid bilayers.

The observation time for each single molecule in solution by FCS-type confocal microscopy is limited to milliseconds, the diffusion time through the confocal volume. For surface-immobilized molecules, the observation time is in principle unlimited: we can take a new image and determine the fluorescence of each molecule at any time. In practice, the observation time will be limited by the photostability of the dye(s) used. Fluorophores undergo *photobleaching* and *blinking*, and transiently or permanently stop emitting fluorescence. Fluorescence emission stops when the fluorophore undergoes a transition to a non-fluorescent dark state, typically a triplet state (see Figure 21.33; Figure 25.20). While in the triplet state, the fluorophore does not contribute to fluorescence. Only after its return to the ground state is the fluorophore available for further excitation-emission cycles and continues to emit fluorescence. The transient loss of fluorescence due to reversible transitions into "dark states" is called blinking. Blinking can be suppressed by addition of *triplet state quenchers*, such as the vitamin E analog Trolox (Box 25.7) that mediates rapid relaxation of triplet states. Molecular oxygen is also a triplet state quencher. Quenching of triplet states by

molecular oxygen generates reactive singlet oxygen that causes photodestruction of the fluorophore and/or covalent modifications, accompanied by a permanent loss of fluorescence. This irreversible process is called photobleaching. Photobleaching can be minimized by using *oxygen scavengers* (Box 25.7), which inevitably leads to more pronounced blinking. The combined use of oxygen scavengers and triplet state quenchers affords typical observation times of tens of seconds up to several minutes.

BOX 25.7: MINIMIZING PHOTOBLEACHING AND BLINKING.

The observation time of a fluorophore is limited by the occurrence of intermittent dark phases (blinking) and irreversible loss of fluorescence due to photobleaching. Oxygen scavenging systems remove molecular oxygen from solutions, and suppress the generation of reactive oxygen species that cause irreversible photodestruction of fluorophores and photobleaching. Two examples of frequently used oxygen scavenging systems employ protocatechuic acid in combination with the enzyme protocatechuate-3,4-dioxygenase, or glucose with glucose oxidase and catalase for oxygen removal (Figure 25.20). Molecular oxygen acts as a triplet state quencher. Its removal not only suppresses photobleaching, but at the same time increases the triplet state lifetime and blinking becomes more pronounced. Blinking has therefore to be suppressed by adding other triplet state quenchers such as β-mercaptoethanol or Trolox, a vitamin E derivative (Figure 25.20). However, high concentrations of β-ME can lead to the population of long-lived dark states.

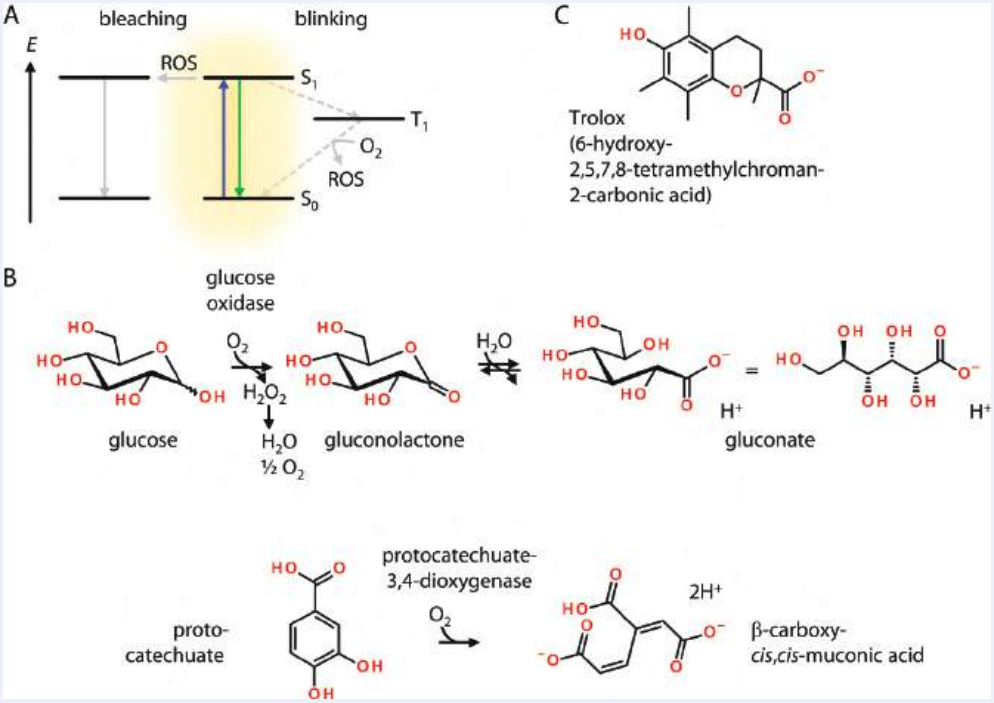


FIGURE 25.20: Photobleaching and blinking limit observation times for single molecules. A: Photobleaching and blinking convert fluorophores into permanent (bleaching) or transient (blinking) dark states. ROS: reactive oxygen species. B: Oxygen scavenging systems suppress photobleaching, but promote blinking. The glucose/glucose oxidase/catalase system removes oxygen by oxidizing glucose to gluconate. Catalase converts the resulting H_2O_2 to H_2O and O_2. Mismatched activities of glucose oxidase and catalase may lead to accumulation of H_2O_2, which can create more reactive oxygen species. The protocatechuate/protocatechuate dioxygenase system removes oxygen by oxidizing protocatechuic acid to β-carboxy-*cis*,*cis*-muconic acid. C: Triplet state quenchers such as Trolox (6-hydroxy-2,5,7,8-tetramethylchroman-2-carbonic acid) suppress blinking by accelerating the relaxation of triplet states.

25.7.2 Why single molecules?

When we measure fluorescence parameters such as intensity, anisotropy, lifetime, or FRET efficiency from solutions over large ensembles of molecules, we obtain values that represent the average over all species that are present in the sample. The average is a meaningful number if the sample is homogeneous, *i.e.* if individual molecules in the ensemble exhibit identical values for the measured parameter or at least similar values narrowly distributed around a mean value (Figure 25.21). The average value does not provide any information on the width of the distribution of the measured parameter or on symmetry or asymmetry of the distribution. For heterogeneous samples with different species that exhibit different fluorescence parameters, the measured value is not meaningful. In this case, the measured value is population-averaged over all molecules in the sample. The average value has very limited meaning on a molecular scale: most likely, this particular value is not represented by a single molecule in our sample. We have discussed this limitation in the context of the FRET efficiency (Section 21.5.10). This limitation is overcome if we determine fluorescence parameters for single molecules individually, without averaging over a large and possibly heterogeneous population. Each single molecule will have one defined value for the respective parameter. By measuring the fluorescence parameter for many single molecules, either simultaneously by imaging a surface on which many molecules are immobilized or one after the other by watching single molecules in solution pass the confocal volume, we directly obtain the distribution of the measured parameter over many molecules in the sample (Figure 25.21). Plots of the frequency with which a particular value is observed are called *probability density functions*. Frequencies determined from experimental data are typically plotted in discrete intervals, as histograms. We have encountered histograms before: photon arrival time histograms to determine fluorescence lifetimes (Figure 21.65 in Section 21.5.9), photon counting histograms to identify species with different brightness (Figure 25.17 in Section 25.6.2), and dwell-time and cumulative dwell time histograms to determine rate constants (Figure 12.3 in Section 12.2).

Single-molecule measurements also eliminate the disadvantage of averaging out temporal changes in the measured parameter. If the molecule of interest can switch between two states with different values for the measured parameter, the transitions of individual molecules in an ensemble from one state to the other are unsynchronized and do not occur in a coordinated manner. These random fluorescence changes are averaged over all molecules at each time point, and the fluorescence intensity from this ensemble of molecules will be constant over time. The only possibility to follow transitions from one state to the other in ensembles is to synchronize the molecules and to make sure that they all start the transition from one state, and at the same time. We will see how we can measure the velocity of transitions in synchronized ensembles in Chapter 28. Often, it is difficult or impossible to trigger a certain transition or reaction, and the molecules cannot be synchronized. If we focus on a single molecule, however, we do not need to synchronize it with others to observe the transition: a single molecule can only be in one of the possible states at any given time. Monitoring the fluorescence intensity of a single molecule over time will therefore provide a fluorescence trace with jumps between the fluorescence levels characteristic of each state (see Section 12.2). According to the principle of ergodicity (Sections 3.1 and 12.2), if given enough time, the single molecule will sample all the states that are represented in a large ensemble of molecules at any given point in time. From the fluorescence traces, we can sum how long this molecule spends in each state, which is a measure for the equilibrium population of this state. Furthermore, we can measure the individual times the molecule spends in a particular state between two switches, the dwell time (Figure 25.21). From these dwell times, we can construct dwell-time histograms, and determine rate constants for the underlying transitions (see Chapter 12). We will revisit these types of information when we discuss single-molecule FRET (Section 25.7.6).

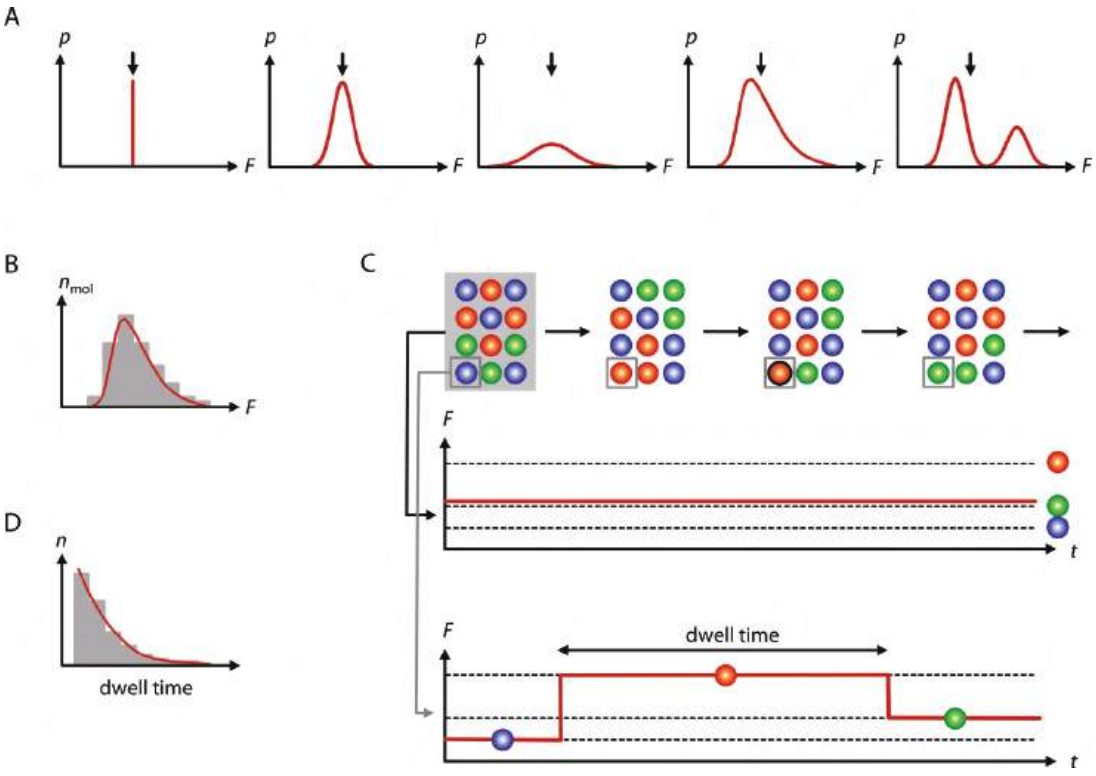


FIGURE 25.21: Information from single-molecule experiments. A: Different distributions $p(F)$ (red) of the observable parameter F and the average value (arrow) obtained from ensemble experiments. B: Single-molecule experiments avoid averaging over the entire population and determine the parameter F individually for a large number of single molecules. The distribution is obtained by plotting histograms. C: Single-molecule experiments avoid temporal averaging and reveal time-dependent changes of an observable for a single molecule over time (bottom). This information is hidden in ensemble experiments (top). From time traces of the observed parameter for a single molecule, dwell times in individual states can be determined. D: Dwell-time histograms provide kinetic information on interconversion of different states (see Chapter 12; Section 25.7.4).

25.7.3 Localization and tracking of single molecules

Imaging single molecules enables us to determine their position. We have discussed before (Section 25.1.3) that a single emitting molecule on the surface appears as a bright spot with a radially symmetric intensity distribution in the fluorescence image, the molecule detection function that corresponds to the Airy disc. The intensity is highest in the center (*i.e.* where the fluorescent molecule is really located), and decays with increasing distance from the center. This intensity distribution is spread over a limited number of pixels of the area detector. Each pixel typically corresponds to a square on the surface of about 0.1 µm by 0.1 µm. Thus, by selecting the highest-intensity pixel, we can only locate the center of the distribution and the position of the emitter with an accuracy in the 0.1 µm range. However, we can exploit the fact that the intensity profile is described by a Gaussian distribution. By describing the intensities measured pixel-wise in the x- and y-directions, we can determine the center of the Gaussian distribution to much higher accuracy, down to 1 nm. This procedure for the localization of single fluorophores is called *fluorescence imaging with one nanometer accuracy* (FIONA, Figure 25.22). The principle of FIONA is also exploited in super-resolution microscopy (Section 25.8).

An alternative to FIONA is MinFlux, a method to localize fluorophores with high accuracy based on the inverse approach: MinFlux probes the sample with a standing wave of excitation light, and exploits the fact that the fluorophore will not emit fluorescence when it is perfectly aligned with the node of this wave (Box 25.8).

Single-molecule microscopy can also be used to track the movement of single molecules over time (Figure 25.22, Figure 25.23). When a single fluorophore is imaged at different time points, its *x*,*y*-coordinates can be determined as a function of time. From this information, *trajectories* of its path can be constructed. This approach has been used to study virus entry into cells and to investigate the lateral diffusion of membrane proteins (see Section 4.1).

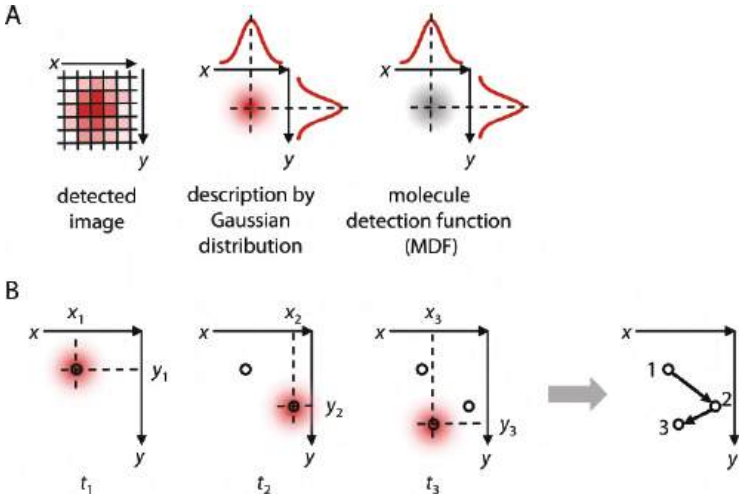


FIGURE 25.22: Localization of single molecules by FIONA and single-molecule tracking.
A: Fluorescence imaging with one nanometer accuracy (FIONA). The continuous intensity distribution from a single emitter (point-spread function, right) is detected by a limited number of pixels on the detector (left). Description of the measured intensity profile by a Gaussian intensity distribution in the *x*- and *y*-directions allows determination of the center of the distribution with nanometer accuracy (center). B: From the position of a single molecule at different time points t_1, t_2, t_3 its trajectory can be constructed (right).

BOX 25.8: MINFLUX AND THE LOCALIZATION OF FLUOROPHORES THROUGH EMISSION MINIMA.

While FIONA localizes fluorophores with high accuracy through the maximum of the emitted fluorescence, MinFlux uses the inverse approach and localizes fluorophores through minima in their emission. To this end, the sample is illuminated with a standing wave of excitation light. Fluorescence from the microscopic sample is imaged with two distinct relative positions of the standing wave with respect to the sample (Figure 25.23). Between these two measurements, the node of the standing wave is moved by a distance L, from one side of the emitter to the other. The emission intensity of the emitter in the two images depends on the light intensity at its position x_E. The intensity profile of the excitation light around the node can be described by a parabola. Using this parabolic function, the position of the emission minimum can be calculated from just two intensities measured. The calculated minimum of the emission corresponds to a node of the standing wave and the position of the emitter. With MinFlux, emitters can be localized with accuracies < 1 nm.

(*Continued*)

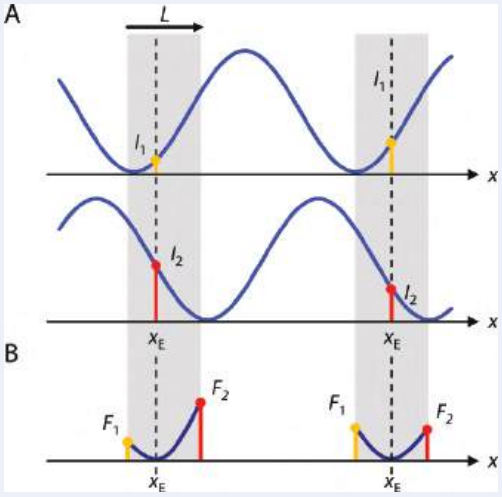


FIGURE 25.23: Localization of single emitters by MinFlux. A: Single emitters localized on the slide surface at the positions x_E are illuminated by a standing wave with a wavelength λ (blue). The excitation intensities, I_1 and I_2, differ for the two different positions of the standing wave relative to the surface. B: The fluorescence intensities F_1 and F_2 of the emitters, measured for the two positions of the standing wave, are described with a harmonic potential (dark blue) to extract the minimum of the emission, which corresponds to the position of the fluorophore. The principle is illustrated for one dimension, but MinFlux can be used for imaging in two dimensions.

25.7.4 Kinetic information from single-molecule microscopy

The fluorescence intensity of a single fluorophore can be measured as a function of time, for example to follow enzymatic reactions in real time. We have discussed the principles of single-molecule kinetics in Chapter 12. Cholesterol oxidase, a bacterial flavoenzyme involved in steroid oxidation and isomerization, uses FAD to oxidize cholesterol to cholest-4-en-3-one. FAD is fluorescent, whereas the reduced form $FADH_2$ is non-fluorescent (Section 21.5.4). Enzymatic turnover can therefore be followed by fluorescence. The time the molecule spends in the on- and off-states, the dwell times (see Section 12.1), can be measured and compiled into a histogram. The mean time a molecule spends in a particular state is its lifetime, which is related to the rate constant of the reaction away from this state (see Chapter 7). From the distribution of dwell times (dwell-time histograms) in the on- and off-states, rate constants for the two processes can therefore be obtained (see Section 12.3).

The first step in the analysis of fluorescence time traces is the assignment of the individual states detected. In simple systems, switching between two states with clearly distinct signals, the state of the molecule at each time point can be assigned by threshold criteria. For more complex systems, involving switches between several states and/or states with small differences in signal, this approach is not possible. In these cases, step-finding algorithms are needed. The most common procedure is the analysis of time traces by hidden Markov modeling (HMM; Box 25.9).

BOX 25.9: HIDDEN MARKOV MODELS.

HMM describes the temporal behavior of a system with a set of different states that exhibit different (fluorescence) signals. These states are connected by single-step processes that are governed by specific rate constants (Figure 25.24). For HMM analysis of fluorescence time traces, the model is described by three matrices, the initiation probability matrix I_p, containing the starting probabilities for states A and B, p_A and p_B,

(*Continued*)

$$I_p = \begin{pmatrix} p_A \\ p_B \end{pmatrix} \qquad \text{eq. 25.33}$$

the transition probability matrix T_p containing the rate constants of interconversion, k_1 and k_{-1},

$$T_p = \begin{pmatrix} k_1 \\ k_{-1} \end{pmatrix} \qquad \text{eq. 25.34}$$

and the emission probability matrix E_p that collects the emission values or signals characteristic of each state, X_A and X_B (Figure 25.24)

$$E_p = \begin{pmatrix} X_A \\ X_B \end{pmatrix} \qquad \text{eq. 25.35}$$

These matrices are then used to identify the states present in the fluorescence time trace, and to determine the most probable path through these states, termed the *Viterbi path*. HMM analysis of single-molecule time traces provides an idealized trajectory, showing only the mean signals of the states populated and the switches between them (Figure 25.24). From the interval between two switching events in the idealized trajectory, the dwell time in a particular state is calculated. Analysis of the resulting dwell-time histograms for states A and B with single-exponential functions then gives the rate constants k_1 and k_{-1} (Section 12.3).

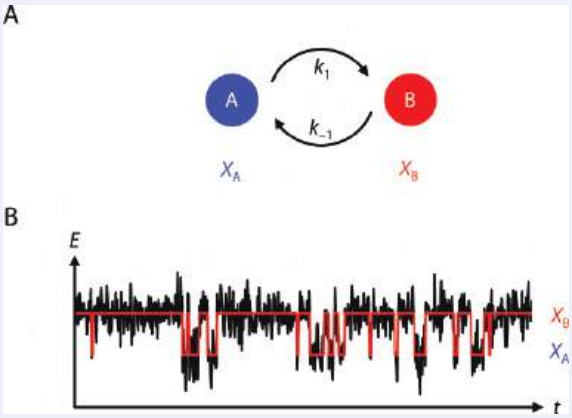


FIGURE 25.24: Analysis of single-molecule time traces by hidden Markov modelling (HMM). A: Kinetic model for hidden Markov analysis of a two-state system with two states, A and B, with signals X_A and X_B, interconverting with rate constants k_1 and k_{-1}. B: In HMM analysis of an experimentally obtained fluorescence time trace (black), the Viterbi path (red), the most probable pathway through the two states is determined using the initiation probability matrix, the transition probability matrix, and the emission probability matrix.

In *protein-induced fluorescence enhancement* (PIFE) experiments, the increase in fluorescence quantum yield upon binding of proteins in the vicinity of a fluorescent dye is used as a probe for protein binding, which can be used to monitor movement of proteins along DNA (Box 25.10). PIFE is sensitive in the distance range of 1–4 nm. Conversely, static quenching of a reporter fluorophore by certain groups can be a useful probe for binding and movement (Box 25.10).

BOX 25.10: PROTEIN-INDUCED FLUORESCENCE ENHANCEMENT (PIFE) AND STATIC QUENCHING AS PROBES FOR BINDING AND MOVEMENT.

Binding of proteins near fluorophores can either lead to an increase in fluorescence (protein-induced fluorescence enhancement, PIFE) or to fluorescence quenching. In both cases, the change in fluorescence is a sensitive probe for the distance between the protein and the respective fluorescent group. Changes in fluorescence intensity report on binding of the protein, and can also report on movement of the protein of interest toward or away from the fluorescent dye (Figure 25.25). PIFE: Hwang & Myong (2014) Chem. Soc. Rev. 43(4): 1221–1229. Quenching: Honda *et al.* (2009) Mol. Cell 35(5): 694–703.

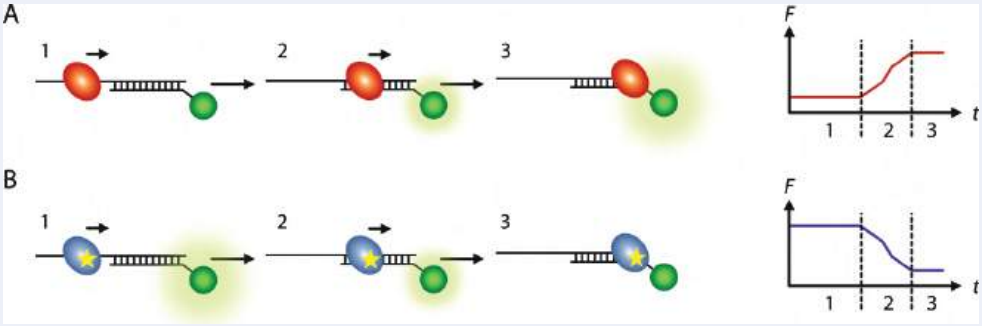


FIGURE 25.25: Fluorescence enhancement and quenching as a probe for binding and movement. A: Binding of a protein close to a fluorescent dye on the DNA may lead to an enhancement of fluorescence. Protein-induced fluorescence enhancement (PIFE) has been used to monitor movement of helicases along DNA. B: Quenching of fluorescence can also serve as a probe for protein movement along DNA. Quenching of fluorescence by the iron-sulfur cluster of the protein XPD has been exploited to measure translocation of XPD along the DNA.

25.7.5 Colocalization of molecules

Multiple color single-molecule fluorescence microscopy can be used to detect individual subunits in multi-component complexes. In *colocalization single-molecule spectroscopy* (CoSMoS), one compound of the complex of interest is immobilized on a surface, and its position is determined by fluorescence microscopy. A second component, labeled with a second dye whose fluorescence is spectrally separated from the fluorescence of the first, is then added, and its position is determined. Superposition of the two images reveals if the two fluorophores colocalize, pointing toward interaction of the partners. In principle, this procedure can be extended to three and more dyes (Figure 25.26). The number of fluorophores that can be used simultaneously is limited by the requirement for spectral separation. The necessity to install separate lasers for excitation of each fluorophore makes multi-color applications rather expensive. CoSMoS provides information on the composition of complexes in equilibrium, and can be used to study complex assembly and disassembly pathways and velocities.

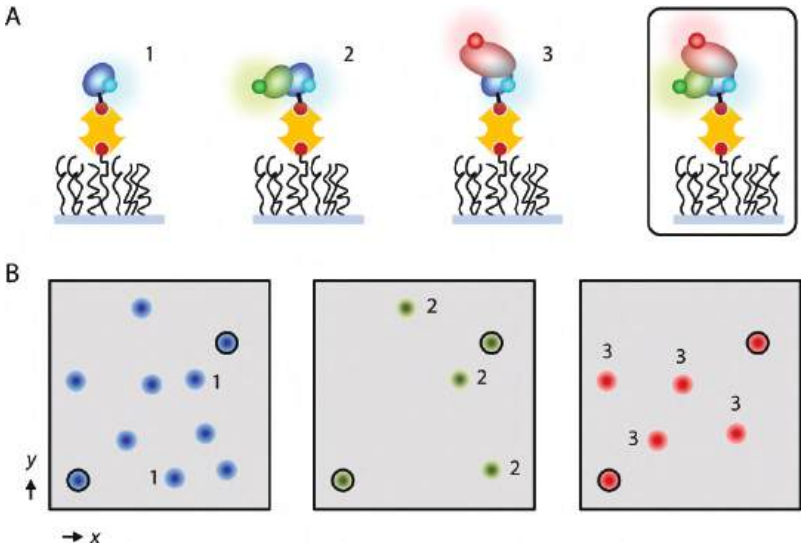


FIGURE 25.26: Colocalization single-molecule spectroscopy (CoSMoS). A: One component of a complex is fluorescently labeled (blue), biotinylated (brown sphere), and immobilized on a streptavidin-functionalized microscope slide. A second and third component are fluorescently labeled with different dyes (green, red). B: The surface is imaged in all three colors, and each labeled component can be localized in the corresponding image. By superimposing the three images, the composition of each complex can be determined. From the ratio of incomplete complexes 1, 2, and 3, and the fully assembled complex (black rim), complex stabilities can be determined. From images at different time points, association and dissociation events can be followed as a function of time.

CoSMoS has been used to study the dynamics of transcription and splicing complexes. The application of CoSMoS is limited to systems that can be assembled *in vitro* from individually purified or synthesized components. Alternatively, the molecule or complex of interest can be pulled down from cell extracts using specific antibodies, and is then surface-immobilized through the antibody for analysis on the single molecule level. This procedure is called *single-molecule pull-down* (SiMPull; Box 25.11).

BOX 25.11: SINGLE-MOLECULE PULL-DOWN (SiMPull).

Single-molecule pull-down combines the principle of pull-down experiments with single-molecule analysis. Proteins or protein complexes can be directly pulled down onto a microscope slide functionalized with an antibody against the protein of interest (Figure 25.27). An extract from cells producing the protein of interest is incubated with the slide surface, and the protein of interest, either alone or in complex with interaction partners, is immobilized on the surface. Unbound proteins are removed by washing. Alternatively, the antibody-bound complex of interest can be immobilized on a protein A-functionalized microscope slide. Protein A recognizes the constant region of immunoglobulins. If the protein has been produced in a fluorescently tagged version, *e.g.* as a fusion with GFP, the surface can be imaged directly. Otherwise, immunolabeling procedures can be employed. Here, a primary antibody against (one of) the immobilized protein(s) is added, and stained by a secondary antibody that is fluorescently labeled. With SiMPull functional complexes from cell extracts may be captured that cannot be reconstituted from individual components *in vitro*. SiMPull of fluorescently labeled proteins and protein complexes enables the determination of complex composition and stoichiometry by analyzing the number of photobleaching steps. Jain *et al.* (2011) Nature 473(7348): 484–488.

(*Continued*)

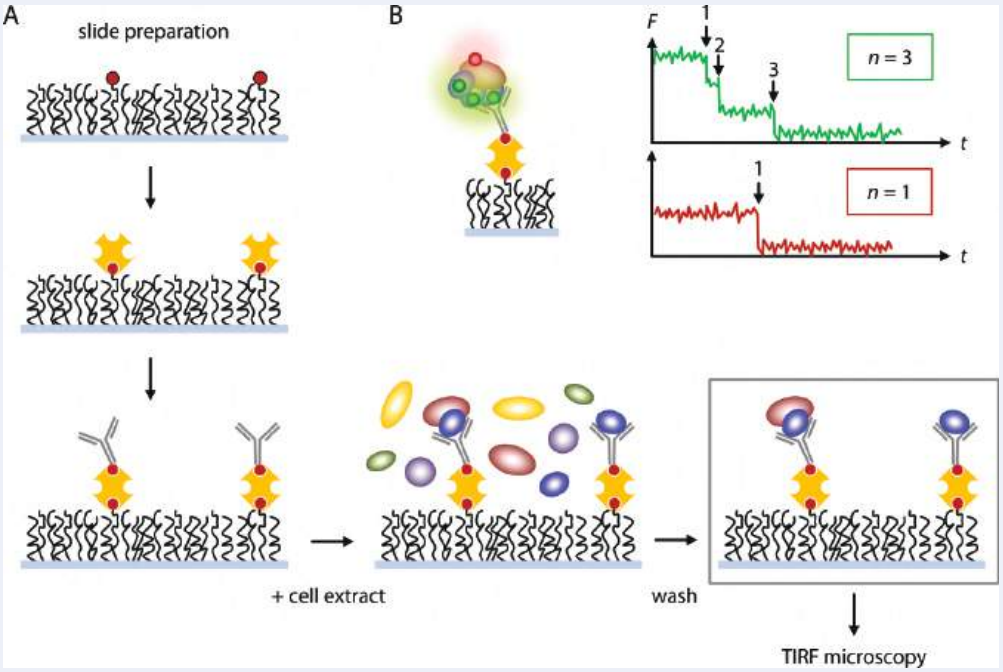


FIGURE 25.27: Single-molecule pull-down (SiMPull). A: Microscope slides are passivated with polyethylene glycol. A small fraction of the polyethylene glycol molecules carries a biotin group (brown sphere) that binds streptavidin (orange). Biotinylated antibodies are immobilized on the streptavidin-functionalized surface. Cell extract is added, and unbound proteins are removed by washing. The antibody captures the protein of interest (blue) and possible interaction partners (brown). The protein (complexes) captured on the surface can be imaged directly if one or both components have been labeled with genetically encoded dyes, such as fluorescent proteins. Alternatively, the surface is immunostained by adding a primary antibody against the interaction partner and a secondary antibody that is fluorescently labeled. B: With SiMPull the composition of isolated complexes can be determined in multi-color experiments. The green fluorescence shows three photobleaching steps, the complex thus contains three of these subunits. The red fluorescence photobleaches in a single step, indicating that only one subunit is part of the complex.

25.7.6 Single-molecule FRET

The presence of two fluorophores enables the measurement of FRET efficiencies and changes of FRET efficiencies due to conformational changes (see Section 21.5.10). Single-molecule FRET microscopy offers the advantage that donor and acceptor emission is measured for single molecules individually instead of averaging intensities over an ensemble. The FRET efficiency from a single molecule therefore directly reports on the inter-dye distance. Histograms of FRET efficiencies for a set of single molecules yield the distribution of FRET efficiencies in the sample, without a prior assumption of the shape of the distribution. When FRET is measured as a function of time, conformational changes can be followed in real time. Single-molecule FRET microscopy can be performed on donor/acceptor-labeled molecules in solution, using confocal microscopy, or on surface-immobilized molecules using confocal scanning, wide-field, or TIRF microscopy. The emission light is separated into donor and acceptor emission by dichroic mirrors (Box 25.1) and detected by separate avalanche photodiodes or imaged onto adjacent areas of a CCD camera. For single-molecule FRET microscopy on freely diffusing molecules, the raw data consist of donor and acceptor fluorescence as a function of time. Spikes of fluorescence stem from a single molecule that passes the confocal volume. Using threshold criteria, single-molecule events can be identified, and the FRET efficiency for each single molecule can be calculated from the donor and acceptor intensities integrated over the duration of the fluorescence burst (Figure 25.28). FRET histograms

reveal the distribution of FRET efficiencies over all molecules observed one after the other, and allow the identification of different conformational states. While single-molecule FRET in solution separates individual molecules in time, molecules are separated in space in single-molecule FRET experiments on surface-immobilized molecules. The raw data consist of pair-wise images, one of the donor fluorescence, one for the acceptor fluorescence, which contain a diffraction-limited spot for each single molecule detected. Integration over these spots allows calculation of the FRET efficiencies for all observed molecules, and for the construction of FRET histograms. From a series of images taken at different time points, the donor acceptor fluorescence and the FRET efficiency can be determined as a function of time. FRET time traces reveal the sequence of different conformational states and the dwell times of the molecule in each state. From dwell-time histograms, rate constants of conformational changes can be determined (Section 25.7.4, see Chapter 12).

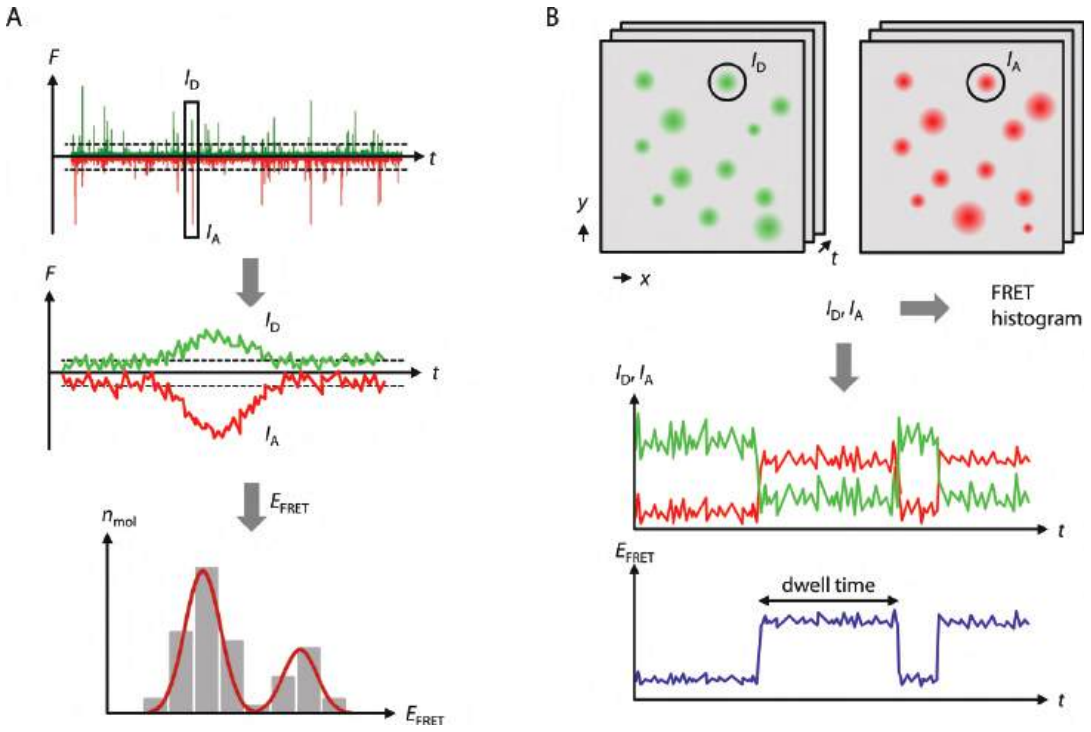


FIGURE 25.28: Single-molecule FRET on molecules in solution or on surfaces. A: Single-molecule FRET in solution by confocal microscopy. Top: raw data: donor (green) and acceptor fluorescence (red, inverted) as a function of time. Each spike of fluorescence is caused by a single molecule diffusing through the confocal volume. The broken lines mark the threshold: bursts with higher intensity are counted as single-molecule events. Middle: Close-up of a single fluorescence burst. The donor and acceptor fluorescence intensities I_D and I_A are obtained by integrating the areas under the curves. From I_D and I_A, the FRET efficiency (E_{FRET}) for the single molecule is calculated. Bottom: FRET efficiencies for all observed molecules are plotted as a FRET histogram. B: Single-molecule FRET of molecules immobilized on a surface. Top: Images of donor (green) and acceptor fluorescence (red) are acquired at different time points. Each molecule appears as a bright spot in the images. The donor and acceptor fluorescence intensities I_D and I_A are determined by integrating the intensities in one spot (black circle). From I_D and I_A, the FRET efficiencies for each molecule can be obtained and plotted as histograms. Middle, bottom: More importantly, the donor and acceptor intensities and thus the FRET efficiency for each single molecule can be calculated as a function of time, revealing conformational changes of the molecule in real time. FRET time traces permit the determination of dwell times in each conformational state (Box 25.9), and the determination of rate constants for the transitions between these states from dwell-time histograms (Chapter 12; Section 25.7.4).

Single-molecule FRET has been widely applied to study protein and nucleic acid conformational changes. The conformational cycle of the translation initiation factor eIF4A, an RNA helicase involved in ATP-dependent unwinding of secondary structures in the 5'-untranslated region of mRNAs during translation initiation, has been investigated by single-molecule FRET confocal and TIRF microscopy. eIF4A adopts three different conformations, termed open, half-open, and closed states. Other translation initiation factors regulate eIF4A activity by modulating the population of these states and the kinetics of their interconversion. Single-molecule FRET can also delineate a series of sequential conformational changes in molecular machines. By introducing donor and acceptor fluorophores in different positions into the DNA topoisomerase gyrase, a cascade of DNA- and nucleotide-induced conformational changes at the beginning of the catalytic cycle of DNA supercoiling has been observed. While most studies detect FRET between a pair of dyes and monitor changes of a single distance, single-molecule FRET experiments can be extended to three colors to monitor two (or more) different distances simultaneously (Box 25.12).

BOX 25.12: THREE-COLOR FRET.

Three-color FRET has been used to investigate the conformational changes in Holliday junctions, four-way DNA helical junctions that are intermediates in homologous recombination (Section 19.5.3.1). The donor fluorophore Cy3 and two acceptor fluorophores, Cy5 and Cy5.5, were attached at the end of three arms of the four-way helical junction, and the Holliday junction was surface-immobilized *via* biotin attached to its fourth arm for single-molecule FRET experiments by TIRF microscopy (Figure 25.29). The Holliday junction can form two X-shaped forms in which two helical arms are co-axially stacked. Two different stacking patterns are possible, with different pairwise stacking of the four helices. Each of these stacking conformers can adopt a parallel or antiparallel form. The stability of stacking conformers is determined by the sequence in the junction region. The measured fluorescence intensities of all three fluorophores were consistent with completely stacked antiparallel Holliday junctions. No intermediates with incomplete stacking were detected. The interconversion of the two antiparallel conformers leads to a decrease in FRET efficiency from donor to one acceptor, and an increase in FRET efficiency between donor and the other acceptor (Figure 25.29). Fluorescence time traces revealed a correlation of the fluorescence increase of one and decrease of the other acceptor (on a time scale of 20 ms), indicating that interconversion of the two antiparallel stacking conformers occurs by a simultaneous and coordinated unstacking of one helix pair and restacking of the alternative pair. Hohng *et al.* (2004) Biophys. J. 87(2): 1328–1337.

(*Continued*)

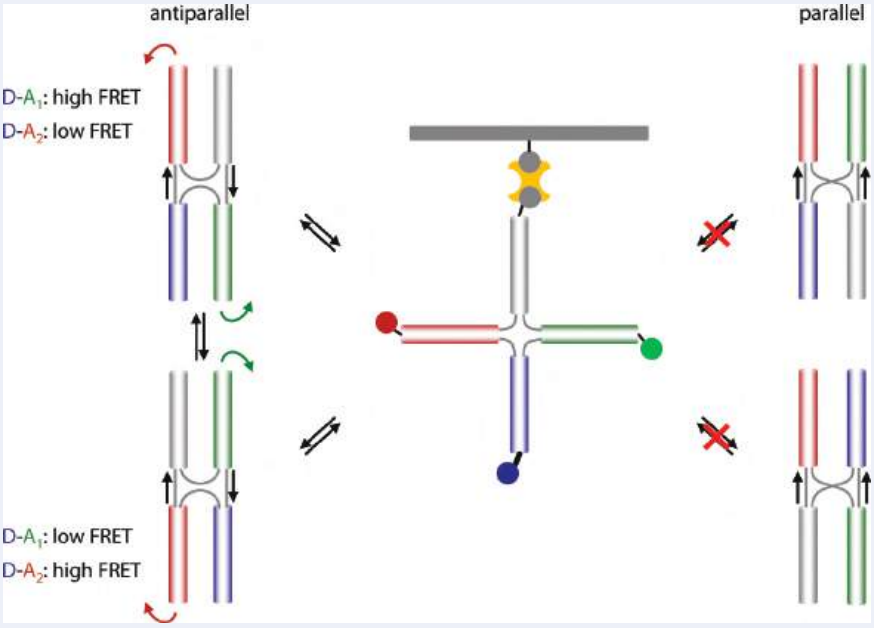


FIGURE 25.29: Three-color single-molecule FRET to study the dynamics of Holliday junctions. A Holliday junction was labeled with donor (blue), acceptor 1 (green), and acceptor 2 (red) on the ends of three helical arms. The fourth helix carried a biotin for surface immobilization. In single-molecule FRET experiments only antiparallel stacking isoforms were detected. During interconversion of these two isoforms by unstacking and restacking of the four helices (colored arrows), the FRET efficiencies between the donor to each of the two acceptors change in opposite directions. The fluorescence intensity of one acceptor increases, the intensity of the second acceptor decreases. In FRET time traces, changes in fluorescence of the two acceptors occurred within a time window of 20 ms, in agreement with a simultaneous movement of the two helices (red, green) during rearrangement.

Distances determined from FRET efficiencies of donor/acceptor-labeled single molecules have a molecular meaning. Provided that the structures of the individual domains or components are known, FRET distances can therefore be used as distance restraints to map the global conformation of multi-domain molecules or to gain insight into the architecture of larger complexes. To map global conformations, donor and acceptor fluorophores are introduced pairwise at various positions in the two parts whose relative orientation is unknown, and the corresponding FRET efficiency is determined in single-molecule experiments (Figure 25.30). To obtain correct FRET efficiencies and correct inter-dye distances, the measured donor and acceptor fluorescence intensities have to be corrected for background fluorescence, for cross-talk of donor fluorescence into the acceptor channel (α) and cross-talk of acceptor fluorescence into the donor channel (β), for different quantum yields of donor and acceptor and different sensitivities of the detectors at donor and acceptor emission wavelengths (γ), and for direct excitation of the acceptor at the excitation wavelength of the donor (δ). The corrected FRET efficiency devoid of these effects is

$$E_{\mathrm{FRET}} = \frac{(1+\beta\gamma\delta)\cdot\left(I_{\mathrm{A}} - \frac{\alpha+\gamma\delta}{1+\beta\gamma\delta}\cdot I_{\mathrm{D}}\right)}{(1+\beta\gamma\delta)\cdot\left(I_{\mathrm{A}} - \frac{\alpha+\gamma\delta}{1+\beta\gamma\delta}\cdot I_{\mathrm{D}}\right) + (\gamma+\gamma\delta)\cdot(I_{\mathrm{D}} - \beta I_{\mathrm{A}})} \qquad \text{eq. 25.36}$$

The corrected FRET efficiencies are then converted into distances (eq. 21.134). Conversion of FRET efficiencies into distances requires the knowledge of the Förster distance for the donor-acceptor pair under the experimental conditions, *i.e.* in the complex studied. Once a set of distances has

been obtained, the two domains of the molecule of interest or the components of a multi-subunit complex are then placed such that the distance restraints from the FRET experiments are fulfilled. FRET-restrained structural modeling takes into account the length of the linker with which the dyes are attached to the molecule of interest and the resulting uncertainties of their positions, and calculates the most probable global conformation from FRET data. Although the procedure is somewhat tedious and requires a large set of donor/acceptor-labeled constructs and calculation effort, nano-positioning by FRET can provide valuable structural information on large complexes whose structures are inherently difficult to determine by high-resolution methods.

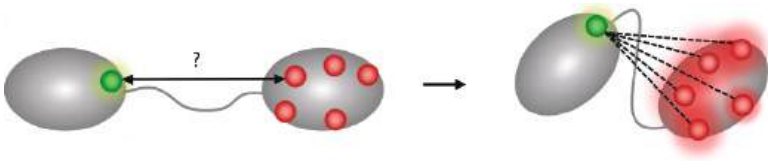


FIGURE 25.30: Principle of FRET-restrained modeling. A two-domain protein of unknown relative domain orientation (left) is labeled with a donor in one domain and an acceptor in different places within the second domain. From the measured pairwise FRET efficiencies, inter-dye distances can be determined (broken lines). The two domains are then placed relative to each other such that the distance constraints are fulfilled.

In standard single-molecule FRET experiments, the donor fluorescence is excited, and donor and acceptor emission are detected. Molecules that do not show acceptor fluorescence might reflect low FRET states, might carry an acceptor fluorophore in a dark state because of blinking or photobleaching, or might not carry an acceptor at all. Integration of a second laser into the microscope enables intermittent direct excitation of the acceptor to probe if the acceptor is still present (Figure 25.31). For excitation with continuous wave lasers, this technique is called alternating laser excitation (ALEX). For pulsed lasers, it is called pulsed interleaved excitation (PIE). In ALEX or PIE experiments, the FRET efficiency E is determined from the donor and acceptor fluorescence intensities upon excitation of the donor:

$$E = \frac{F_{\mathrm{Dex}}^{\mathrm{Aem}}}{F_{\mathrm{Dex}}^{\mathrm{Dem}} + F_{\mathrm{Dex}}^{\mathrm{Aem}}} \qquad \text{eq. 25.37}$$

The subscript Dex denotes donor excitation, the superscripts Dem and Aem denote donor and acceptor emission, respectively. Note that eq. 25.37 does not take into account corrections for instrument non-idealities (see eq. 25.36).

In addition to the FRET efficiency, the stoichiometry S of the fluorescent dyes can be determined by relating the sum of donor and acceptor fluorescence intensities upon donor excitation and acceptor excitation:

$$S = \frac{F_{\mathrm{Dex}}^{\mathrm{D+A}}}{F_{\mathrm{Dex}}^{\mathrm{D+A}} + F_{\mathrm{Aex}}^{\mathrm{D+A}}} \qquad \text{eq. 25.38}$$

For a molecule that carries donor and acceptor, the sums of donor and acceptor fluorescence ($F^{\mathrm{D+A}}$) will be equal for donor and acceptor excitation, and a value of $S = 0.5$ is obtained. For molecules that carry only a donor, F_{Aex} is small, and $S \rightarrow 1$, whereas for molecules that carry only an acceptor, F_{Dex} is much smaller than F_{Aex}, and $S \rightarrow 0$. FRET efficiency E and stoichiometry S can be plotted in two-dimensional histograms (Figure 25.31). From these histograms, it is evident that molecules with the same FRET efficiency that carry different labels can now clearly be distinguished. ALEX and PIE afford spectral sorting of species that carry both the donor and the acceptor from other (unwanted) species that contain only one of the dyes. Donor only- and acceptor-only peaks in FRET histograms can be suppressed, and the subsequent data analysis is restricted to donor/acceptor-labeled molecules. This approach is particularly important for molecules where complete labeling cannot be achieved.

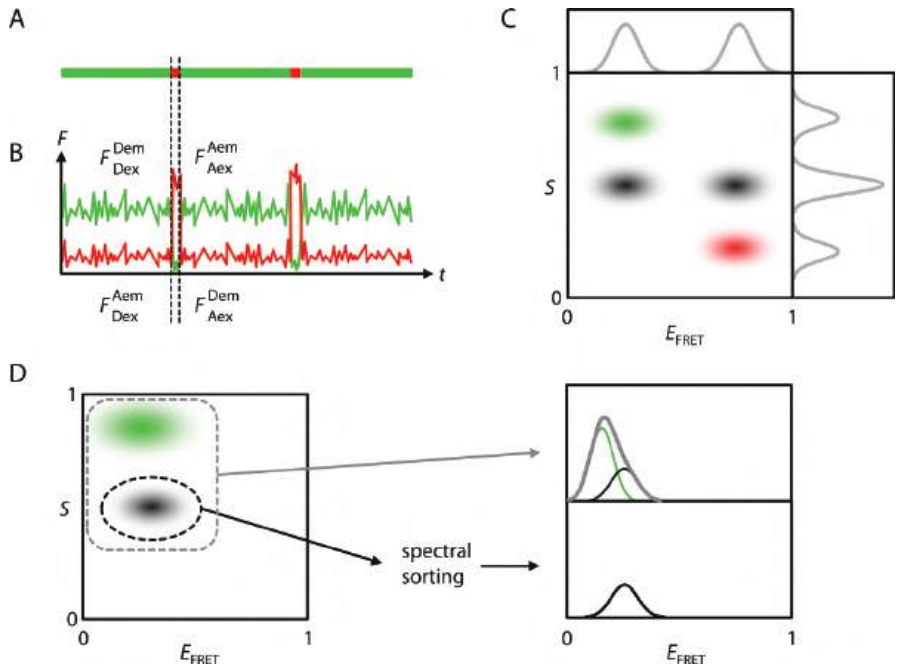


FIGURE 25.31: ALEX/PIE and spectral sorting. A: Illumination scheme. In ALEX/PIE, the sample is illuminated with light to excite donor fluorescence (green) and intermittently with light of a different wavelength to excite the acceptor fluorescence directly (red). B: While the donor fluorescence is excited, the donor and acceptor fluorescence intensities, F^{Dem}_{Dex} and F^{Aem}_{Dex}, report on FRET. In the intervals when the acceptor fluorescence is excited, the donor fluorescence F^{Dem}_{Aex} drops to background level, and F^{Aem}_{Aex} reports on the presence of the acceptor. From the individual intensities, FRET efficiency E (eq. 25.37) and stoichiometry S (eq. 25.38) can be calculated. C: Two-dimensional histogram of FRET efficiency E and stoichiometry S for a donor/acceptor-labeled molecule. 25% of the molecules carry only a donor (green), 50% carry both donor and acceptor and occur in a high- and low-FRET state (black), and 25% carry only the acceptor (red). The individual E and S histograms are shown on top and on the right, respectively. D: Spectral sorting. E-S-histogram for a sample with a low FRET donor/acceptor-labeled molecule (black) that contains a large fraction of molecules labeled only with the donor (green). Without spectral sorting, the FRET histogram (top right, gray) is dominated by the FRET = 0 peak from the donor only molecules. The donor/acceptor-labeled molecules cause a shoulder, and their FRET efficiency distribution (black) cannot be extracted reliably. The FRET histogram constructed from only those molecules that carry both dyes and give rise to FRET (bottom right) reveals a clear FRET distribution with a well-defined maximum.

25.8 SUPER-RESOLUTION MICROSCOPY

We have seen in Section 25.1.3 that focusing excitation light by a lens does not lead to a tiny, infinitesimally small focus, but to a spot of finite dimensions because of diffraction effects. The diffraction-limited resolution of light microscopy is a physical law and cannot be overcome. Nevertheless, developments have been made to circumvent these limitations, and to obtain higher resolution.

Moderate increases in optical resolution are achieved by confocal microscopy using two-photon excitation (Box 25.4), by *structured illumination microscopy* where the sample is illuminated in periodic patterns, or by *4Pi microscopy* that uses two opposing objectives for excitation and emission to increase the apparent numerical aperture of the lens system. More drastic reductions of the point-spread function and increases in resolution can be achieved by introducing sub-diffraction-limit features into the excitation light. For example, the size of the point-spread function can be reduced by taking advantage of stimulated emission (Figure 25.32). In a first step, the excitation light is focused into a diffraction-limited spot, leading to excitation of molecules within this area. In a second step, an annular (ring-shaped) laser beam is used to illuminate the rim of the excitation spot. In the ring illuminated by this second laser, stimulated emission occurs, *i.e.* the laser light induces transitions of the fluorophores from the excited state back to the ground state. This causes a depletion of molecules

in the excited state at the rim of the spot (*stimulated emission depletion, STED*). As a result, a central spot of excited molecules is left that is now smaller than the diffraction limit, effectively reducing the point-spread function.

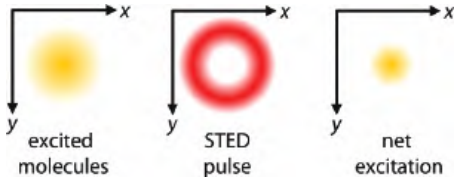


FIGURE 25.32: Stimulated emission depletion (STED) microscopy. In STED microscopy, a diffraction-limited spot of the sample is illuminated and molecules within this spot are excited (left, orange). With a second high-power laser (STED laser) with an annular intensity profile (red, center), molecules at the rim of the focal spot are illuminated and undergo stimulated emission from the excited to the ground state. As a result, a spot of excited molecules remains that is smaller than the diffraction-limited spot (orange, right). With this smaller focal spot, higher resolution imaging is achieved.

Because of the high laser powers needed for stimulated emission depletion, the application of STED is limited to highly photostable fluorophores. In super-resolution microscopy based on the *reversible saturable optically linear fluorescence transitions* (RESOLFT) technique, the diameter of the diffraction-limited spot is reduced by depletion of fluorophores that can be switched reversibly between on- and off-states. This switching can be achieved with lower laser power than stimulated emission depletion, and at the same time expands the range of suitable fluorophores. With STED and RESOLFT microscopy, lateral resolutions (in the *xy*-plane) of 30–50 nm are achieved (Figure 25.33).

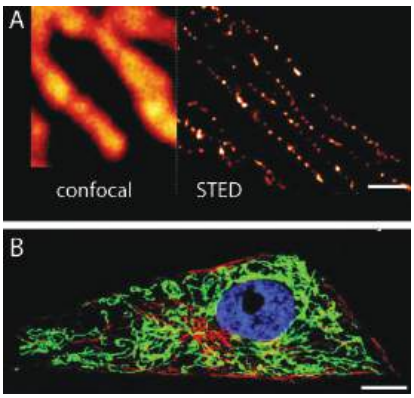


FIGURE 25.33: STED microscopy. A: Comparison of STED and confocal images of mitochondria. The TOM complex was immunostained using fluorescently labeled antibodies. The white bar corresponds to 500 nm. B: STED image of the mitochondrial network of a Ptk2 kangaroo rat cell. The TOM complex (green) and the microtubules (red) were immunostained using fluorescently labeled antibodies, the DNA in the nucleus (blue) is stained with DAPI. The white bar corresponds to 10 µm. Adapted from Schmidt, R., Wurm, C.A., Punge, A., Egner, A., Jacobs, S., Hell, S.W. (2009) Mitochondrial cristae revealed with focused light, Nano Lett. 9(6) 2508–2510 with permission. Copyright 2009, American Chemical Society.

A conceptually different set of high-resolution imaging approaches exploits the defined shape of the point-spread and molecule detection functions for a single emitter and locates the emitter with nanometer accuracy from the center of the Gaussian intensity profile (FIONA, Section 25.7.3). The location of a single fluorescent molecule can be determined more accurately than the Abbe limit predicts. However, imaging of cells or sub-cellular structures requires the accurate localization of a large number of fluorophores. Each of these fluorophores causes a fluorescence spot on the detector. If these fluorescence spots overlap, localization of individual molecules becomes difficult or impossible. Under certain circumstances, the FIONA localization principle can also be applied for imaging

of such large ensembles of fluorophores. All we need to do is to selectively activate a subset of the fluorophores present that are separated by a distance exceeding the diffraction limit. We can then determine the position of these fluorophores accurately. In the next round, we activate a different subset of fluorophores, and determine their positions. After several rounds, we will statistically have determined the position of each fluorophore at least once. A complete image is constructed from all partial images (Figure 25.34). This principle is the basis of *(fluorescence) photoactivated localization microscopy* (PALM or FPALM). While the fluorophores are switched on and off by illumination in (F)PALM, a related type of microscopy, *stochastic optical reconstruction microscopy* (STORM; Figure 25.35), relies on stochastic switching between on and off states of the fluorophore to limit the number of fluorophores in each partial image.

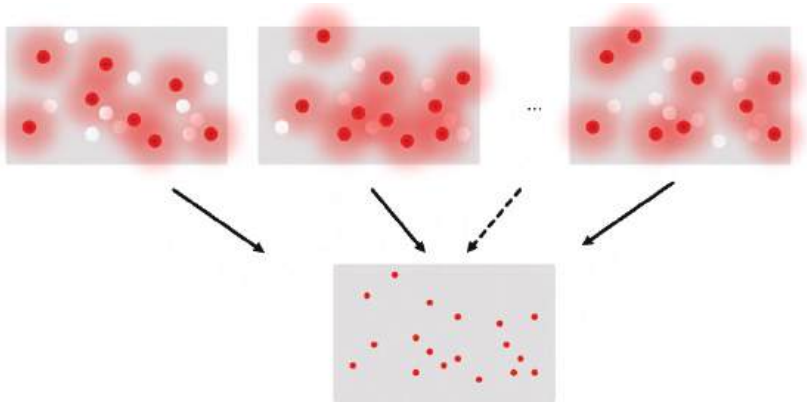


FIGURE 25.34: Principle of PALM and STORM. A fluorescent image is the sum of a large number of single emitters. Each emitter gives rise to an Airy disc in the image, but their Airy discs overlap and lead to a blurred image. PALM and STORM use fluorescent dyes that can be switched on and off by light (PALM) or that stochastically switch between dark and emitting states (STORM). In both cases, only a small subset of emitters is "on" at any given point in time. The Airy discs from these few emitters are now clearly separated in the image (top), and their centers can be determined with nanometer accuracy (FIONA, see Figure 25.22). From a series of images with only a subset of dyes activated, the positions of all emitters can be determined, and the entire image can be constructed (bottom).

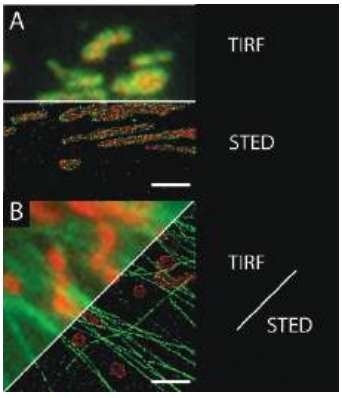


FIGURE 25.35: STORM microscopy. A: Comparison of STORM and TIRF images of African green monkey kidney cells. Green: outer mitochondrial membrane protein Tom20, red: inner mitochondrial membrane protein ATP synthase. Proteins were immunostained using fluorescently labeled secondary antibodies. B: STORM and TIRF image of Tom20 (red) and microtubules (green). White bars correspond to 1 µm. Adapted with permission from Tam, J. & Merino, D. (2015) Stochastic optical reconstruction microscopy (STORM) in comparison with stimulated emission depletion (STED) and other imaging methods. J. Neurochem. 135(4), 643–658. Copyright 2015, John Wiley and Sons.

Super-resolution microscopy can also be achieved by high-accuracy localization of individual emitters through the MinFlux principle (see Section 25.7.3, Box 25.8). Here, fluorophores are detected through minimal fluorescence emission when they are aligned with the node of a standing wave used for excitation.

All super-resolution methods treated in this Section strive to increase resolution by circumventing Abbe's law to localize single emitters with high precision. In *expansion microscopy*, a complementary approach is used. Here, the resolution is increased by increasing the dimensions of the sample through physical manipulation before the imaging process (Figure 25.36). Expansion of the sample increases the distance between neighboring emitting molecules, and thus of their point-spread function, enabling their imaging as separate emitters. Expansion microscopy can be combined with all super-resolution methods introduced in this Section to increase resolution even further.

For the expansion process, the specimen is fixed, the protein of interest is stained, typically by immunostaining with fluorescently labeled antibodies, and the specimen is embedded into a hydrogel. After covalent crosslinking of the antibody and hence the fluorophores to the molecules of the gel matrix, the proteins are proteolytically cleaved. In the final step, the hydrogel is hydrated to allow it to swell. During this process the distance between the fluorophores increases by 2–4-fold, depending on the gel matrix. It is also possible to perform the staining as the last step, after gel expansion. If the protein of interest is fused to a fluorescent protein for imaging, no staining is required.

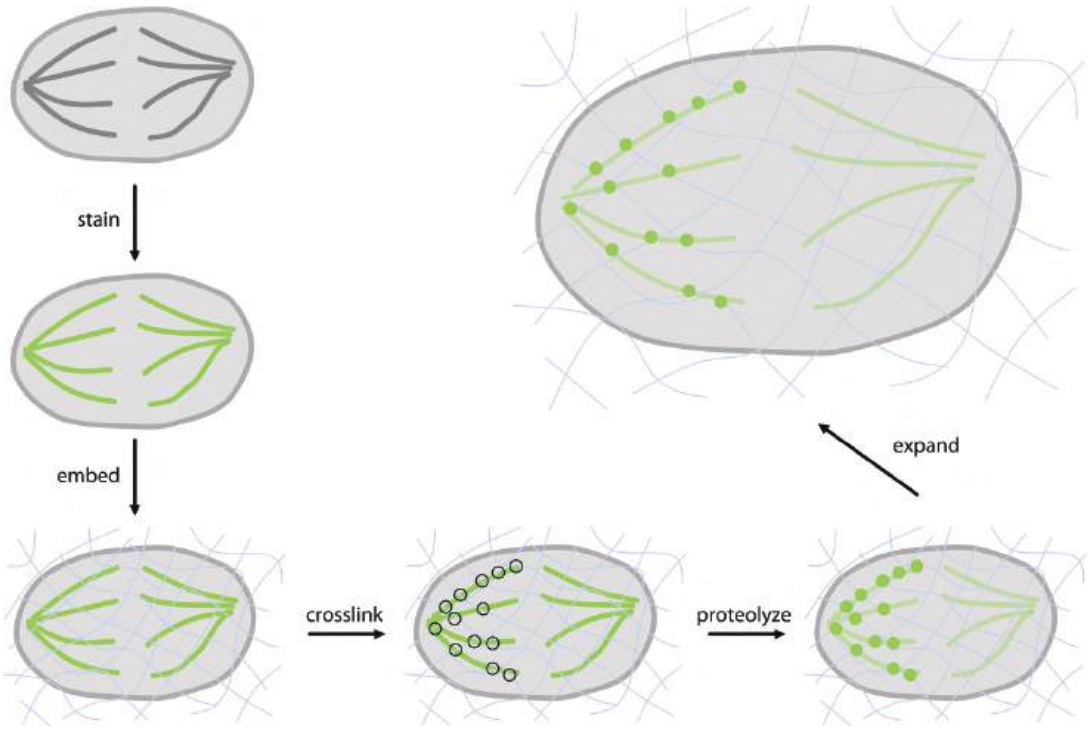


FIGURE 25.36: Expansion microscopy. The specimen is stained (green), embedded into a hydrogel (light blue mesh), and covalently crosslinked to the gel matrix (black circles). After protease treatment, swelling of the hydrogel is initiated, which leads to an expansion of the specimen and an increase in inter-fluorophore distance (green dots) by a factor of approx. 2–4 in all three dimensions.

QUESTIONS

25.1 You need to expand a laser beam from $d_1 = 0.2$ mm to $d_2 = 5$ cm to fully illuminate the back-focal plane of your objective. The microscope has two empty lens holders with a distance of 26 cm. What lenses can you use?

25.2 A rapidly diffusing molecule has a very short residence time in the confocal volume of an FCS instrument. How do you have to change the size of the confocal volume to be able to measure the diffusion time? Does this change require a widening or a focusing of the excitation beam?

25.3 In an FCS experiment, the following fluorescence intensities are measured as a function of time.

t (ns)	intensity
0.1	1
0.2	1
0.3	2
0.4	4
0.5	1
0.6	1
0.7	1
0.8	1
0.9	0
1	0

Calculate the autocorrelation curve for $\tau = 0$ ns to $\tau = 1$ ns (in 0.1 ns steps). What is the diffusion time of the molecule through the confocal volume?

25.4 By FCS, the diffusion time of a double-stranded DNA is measured as $\tau_{diff,DNA} = 3$ ns. In the presence of saturating concentrations of a restriction enzyme, the diffusion time increases to $\tau_{diff,complex} = 10$ ns. At a concentration of $c_{DNA} = 100$ nM and an enzyme concentration $c_{enzyme} = 500$ nM, an apparent diffusion time of 8 ns is measured. What is the K_d value of the enzyme-DNA complex?

25.5 Why do fluorescence bursts from single-molecule FRET by confocal microscopy show a rather wide distribution of intensities and durations?

25.6 How can you use fluorescence cross-correlation (FCCS) to test if a FRET sample contains species that carry both the donor and the acceptor fluorophore?

25.7 In a total internal reflection fluorescence (TIRF) microscope the evanescent field is created at the glass-water interface ($n_1 = n_{glass} = 1.51$ and $n_2 = n_{water} = 1.33$). Calculate the critical angle θ_c for total internal reflection. What is the penetration depth d of the evanescent field into the sample (water) at θ_c for excitation light with $\lambda = 280$ nm, $\lambda = 495$ nm, and $\lambda = 800$ nm. What is d if the beam hits the interface with $\theta = 70°$ and $\theta = 80°$?

25.8 From a double-exponential dwell-time histogram, the rate constants $k_1 = 1\ s^{-1}$ and $k_2 = 5\ s^{-1}$ are extracted. The amplitude of the fast phase of the dwell-time histogram is 60%. What is the fraction of molecules reacting in the fast reaction?

25.9 Compare the information from dwell-time histograms obtained by single-molecule microscopy with the information from a histogram of photon arrival times in time-correlated single photon counting.

REFERENCES

Andreou, A. Z. and Klostermeier, D. (2014) eIF4B and eIF4G jointly stimulate eIF4A ATPase and unwinding activities by modulation of the eIF4A conformational cycle. *J Mol Biol* **426**(1): 51–61.
· identification of three conformational states of the translation initiation factor eIF4A by confocal microscopy

Balzarotti, F., Eilers, Y., Gwosch, K. C., Gynnå, A. H., Westphal, V., …, Hell, S. W. (2017) Nanometer resolution imaging and tracking of fluorescent molecules with minimal photon fluxes. *Science* **355**(6325): 606–612.
· original paper describing super-resolution microscopy with MinFlux

Baudendistel, N., Muller, G., Waldeck, W., Angel, P. and Langowski, J. (2005) Two-hybrid fluorescence cross-correlation spectroscopy detects protein-protein interactions *in vivo*. *Chemphyschem* **6**(5): 984–990.
· *in vivo* study of the interaction between fluorescently labeled transcription factors jun and fos by FCCS

Blanco, M. and Walter, N. G. (2010) Analysis of complex single-molecule FRET time trajectories. *Methods Enzymol* **472**: 153–178.
· review describing different strategies in trajectory analysis

Chakraborty, A., Krause, L. and Klostermeier, D. (2022) Determination of rate constants for conformational changes of RNA helicases by single-molecule FRET TIRF microscopy. *Methods* **204**: 428–441.
· review describing the analysis of trajectories to determine rate constants

Chozinski, T. J., Halpern, A. R., Okawa, H., Kim, H.-J., Tremel, G. J., Wong, R. O. L. and Vaughan, J. C. (2016) Expansion microscopy with conventional antibodies and fluorescent proteins. *Nat Methods* **13**(6): 485–488.
· review on expansion microscopy

Deguchi, T., Iwanski, M. K., Schentarra, E.-M., Heidebrech, C., Schmid, L., …, Ries, J. (2023) Direct observation of motor protein stepping in living cells using MINFLUX. *Science* **379**(6636): 1010–1015.
· time-resolved MinFlux

Dertinger, T., Pacheco, V., von der Hocht, I., Hartmann, R., Gregor, I. and Enderlein, J. (2007) Two-focus fluorescence correlation spectroscopy: A new tool for accurate and absolute diffusion measurements. *Chemphyschem* **8**(3): 433–443.
· accurate determination of diffusion coefficients by dual-focus FCS

Dorlich, R. M., Chen, Q., Hedde, P. N., Schuster, V., Hippler, M., Wesslowski, J., Davidson, G. and Nienhaus, G. U. (2015) Dual-color dual-focus line-scanning FCS for quantitative analysis of receptor-ligand interactions in living specimens. *Sci Rep* **5**: 10149.
· determination of ligand affinities of membrane receptors within the membrane of living cells by dual-focus FCS and line-scanning

Egner, A. and Hell, S. W. (2005) Fluorescence microscopy with super-resolved optical sections. *Trends Cell Biol* **15**(4): 207–215.
· review article on 4Pi super-resolution microscopy

Friedman, L. J., Chung, J. and Gelles, J. (2006) Viewing dynamic assembly of molecular complexes by multi-wavelength single-molecule fluorescence. *Biophys J* **91**(3): 1023–1031.
· principle of multi-color detection that is the basis for colocalization single-molecule spectroscopy (CoSMoS)

Gell, C., Brockwell, D. and Smith, A. (2006) *Handbook of Single-Molecule Fluorescence Spectroscopy*. Oxford, Oxford University Press.
· single-molecule microscopy

Gubaev, A. and Klostermeier, D. (2014) The mechanism of negative DNA supercoiling: A cascade of DNA-induced conformational changes prepares gyrase for strand passage. *DNA Repair* **16**: 23–34.
· review article summarizing single-molecule FRET studies that identified a series of conformational changes in the catalytic cycle of DNA gyrase

Harms, U., Andreou, A. Z., Gubaev, A. and Klostermeier, D. (2014) eIF4B, eIF4G and RNA regulate eIF4A activity in translation initiation by modulating the eIF4A conformational cycle. *Nucleic Acids Res* **42**(12): 7911–7922.
· investigation of the kinetics of the conformational cycle of the translation initiation factor eIF4A and its regulation by interaction partners by single-molecule FRET using TIRF microscopy

Hell, S. W. and Kroug, M. (1995) Ground-state depletion fluorescence microscopy, a concept for breaking the diffraction resolution limit. *Appl Phys B* **60**: 495–497.
· original description of the concept of stimulated emission depletion (STED) microscopy

Hohng, S., Joo, C. and Ha, T. (2004) Single-molecule three-color FRET. *Biophys J* **87**(2): 1328–1337.
· three-color FRET study that monitors two distance vectors in Holliday junctions simultaneously

Honda, M., Park, J., Pugh, R. A., Ha, T. and Spies, M. (2009) Single-molecule analysis reveals differential effect of ssDNA-binding proteins on DNA translocation by XPD helicase. *Mol Cell* **35**(5): 694–703.
· measurement of translocation by the DNA helicase XPD along DNA by fluorescence quenching

Huang, B., Bates, M. and Zhuang, X. (2009) Super-resolution fluorescence microscopy. *Annu Rev Biochem* **78**: 993–1016.
· review article comparing different super-resolution microscopy techniques

Hwang, H. and Myong, S. (2014) Protein induced fluorescence enhancement (PIFE) for probing protein-nucleic acid interactions. *Chem Soc Rev* **43**(4): 1221–1229.
· review article on applications of protein-induced fluorescence enhancement (PIFE)

Jain, A., Liu, R., Ramani, B., Arauz, E., Ishitsuka, Y., Ragunathan, K., Park, J., Chen, J., Xiang, Y. K. and Ha, T. (2011) Probing cellular protein complexes using single-molecule pull-down. *Nature* **473**(7348): 484–488.
· original description of single-molecule pull-down experiments

Klar, T. A., Jakobs, S., Dyba, M., Egner, A. and Hell, S. W. (2000) Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proc Natl Acad Sci U S A* **97**(15): 8206–8210.
· first achievement of super-resolution by stimulated emission depletion (STED) microscopy

Lee, T. H. (2009) Extracting kinetics information from single-molecule fluorescence resonance energy transfer data using hidden Markov Models. *J Phys Chem* **113**(33): 11535–11542.
· application of hidden Markov modeling

Lommerse, P. H., Snaar-Jagalska, B. E., Spaink, H. P. and Schmidt, T. (2005) Single-molecule diffusion measurements of H-Ras at the plasma membrane of live cells reveal microdomain localization upon activation. *J Cell Sci* **118**(Pt 9): 1799–1809.
· single-molecule tracking to investigate lateral diffusion of the membrane-associated protein Ras

McKinney, S. A., Joo, C. and Ha, T. (2006) Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys J* **91**(5): 1941–1951.
· application of hidden Markov modeling

Ohrt, T., Mutze, J., Staroske, W., Weinmann, L., Hock, J., …, Schwille, P. (2008) Fluorescence correlation spectroscopy and fluorescence cross-correlation spectroscopy reveal the cytoplasmic origination of loaded nuclear RISC *in vivo* in human cells. *Nucleic Acids Res* **36**(20): 6439–6449.
· *in vivo* FCS/FCCS study on the RNA-induced silencing complex that identified cytoplasmic and nuclear forms with different diffusion coefficients

Okumus, B., Arslan, S., Fengler, S. M., Myong, S. and Ha, T. (2009) Single molecule nanocontainers made porous using a bacterial toxin. *J Am Chem Soc* **131**(41): 14844–14849.
· lipid vesicles containing the pore-forming toxin hemolysin as nano-containers for single-molecule experiments

Rasnik, I., McKinney, S. A. and Ha, T. (2006) Nonblinking and long-lasting single-molecule fluorescence imaging. *Nat Methods* **3**(11): 891–893.
· validation of the triplet state quencher Trolox to minimize blinking

Schaaf, M. J., Koopmans, W. J., Meckel, T., van Noort, J., Snaar-Jagalska, B. E., Schmidt, T. S. and Spaink, H. P. (2009) Single-molecule microscopy reveals membrane microdomain organization of cells in a living vertebrate. *Biophys J* **97**(4): 1206–1214.
· single-molecule tracking of Ras in membranes of living zebrafish embryos

Schmidt, R., Wurm, C. A., Punge, A., Egner, A., Jakobs, S. and Hell, S. W. (2009) Mitochondrial cristae revealed with focused light. *Nano Lett* **9**(6): 2508–2510.
· imaging of mitochondrial fine-structure using stimulated emission depletion (STED) microscopy

Seisenberger, G., Ried, M. U., Endress, T., Buning, H., Hallek, M. and Brauchle, C. (2001) Real-time single-molecule imaging of the infection pathway of an adeno-associated virus. *Science* **294**(5548): 1929–1932.
· single-virus tracking to study virus entry pathways into cells

Selvin, M. and Ha, T. (2008) *Single-molecule Techniques – A Laboratory Manual.* New York, Cold Spring Harbour Press.
· an overview about single-molecule techniques, including single-molecule fluorescence microscopy, FRET and FCS, with a focus on practical aspects and protocols.

Sustarsic, M. and Kapanidis, A. N. (2015) Taking the ruler to the jungle: Single-molecule FRET for understanding biomolecular structure and dynamics in live cells. *Curr Opin Struct Biol* **34**: 52–59.
· review article highlighting advances in applications of single-molecule FRET *in vivo*

Tam, J. and Merino, D. (2015) Stochastic optical reconstruction microscopy (STORM) in comparison with stimulated emission depletion (STED) and other imaging methods. *J Neurochem* **135**(4): 643–658.
· review article comparing different super-resolution and conventional microscopy techniques

Tillberg, P. W., Chen, F., Piatkevich, K. D., Zhao, Y., Yu, C. J., …, Boyden, E. S. (2016) Protein-retention expansion microscopy of cells and tissues labeled using standard fluorescent proteins and antibodies. *Nat Biotechnol* **34**(9): 987–992.
· review summarizing experimental strategies in expansion microscopy

Tokunaga, M., Imamoto, N. and Sakata-Sogawa, K. (2008) Highly inclined thin illumination enables clear single-molecule imaging in cells. *Nat Methods* **5**(2): 159–161.
· description of highly inclined and laminated optical sheet microscopy (HILO) and its application in *in vivo* imaging

Wachsmuth, M., Conrad, C., Bulkescher, J., Koch, B., Mahen, R., Isokane, M., Pepperkok, R. and Ellenberg, J. (2015) High-throughput fluorescence correlation spectroscopy enables analysis of proteome dynamics in living cells. *Nat Biotechnol* **33**(4): 384–389.
· high-throughput FCS (and FCCS) format to study the dynamics of large numbers of proteins in living cells by time-lapse acquisition of FCS data at specific subcellular locations

Weiss, K. and Enderlein, J. (2012) Lipid diffusion within black lipid membranes measured with dual-focus fluorescence correlation spectroscopy. *Chemphyschem* **13**(4): 990–1000.
· application of dual-focus FCS to lipid diffusion in black lipid membranes

Wolff, J. O., Scheiderer, L., Engelhard, T., Engelhardt, J., Matthias, J. and Hell, S. W. (2023) MINFLUX dissects the unimpeded walking of kinesin-1. *Science* **379**(6636): 1004–1010.
· time-resolved MinFlux

Yildiz, A. and Selvin, P. R. (2005) Fluorescence imaging with one nanometer accuracy: Application to molecular motors. *Acc Chem Res* **38**(7): 574–582.
· description of fluorescence imaging with one nanometer accuracy (FIONA)

26

Electron microscopy

Electron microscopy (EM) is an example of *charged-particle microscopy*, where, instead of light, electrons are focused by magnetic lenses onto the sample. These electrons interact with and are scattered by the sample, leading to changes in their direction of travel (Figure 26.1). The interactions of electrons with the sample will be treated in more detail in Section 26.2. After their interaction with the object, the electrons are manipulated by magnetic lenses for analysis. Depending on the kind of analysis, different EM techniques can be distinguished that deliver either surface or internal information on the object: *Scanning electron microscopy* (SEM) uses information from electrons reflected by the sample and provides information on the sample surface, while *transmission electron microscopy* (TEM; Figure 26.1) uses transmitted electrons that pass through the sample and provides information on the inner structure of the specimen.

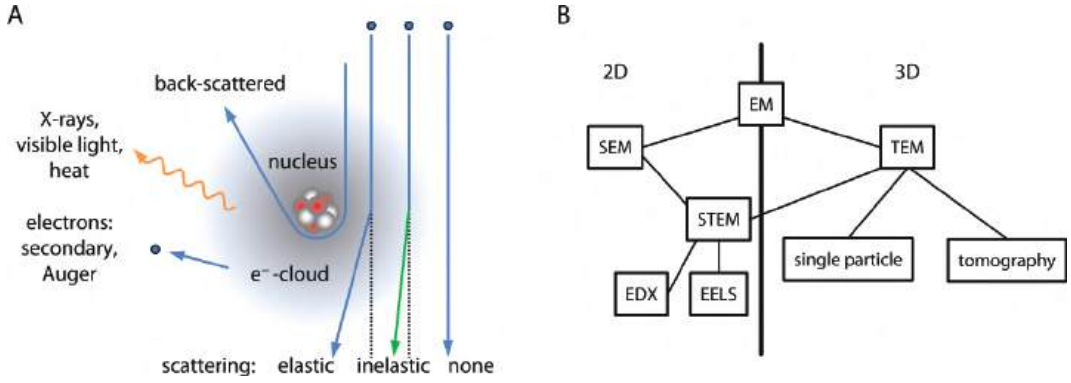


FIGURE 26.1: Techniques enabled by electron-matter interactions. A: Elastic and inelastic scattering. Most electrons (blue spheres and arrows) impinging on a thin object do not interact at all with it, only some of them are scattered. Elastic scattering of electrons by the nucleus of an atom changes the direction of travel. Electrons with large scattering angles (back-scattering) are analyzed in SEM. Inelastic scattering changes not only the direction of travel of electrons, but also their energy, leading to an increase in wavelength (green arrow). B: Main electron microscopy techniques. Electron microscopy is based on the structural information contained in elastically scattered electrons. TEM reveals the internal structure of the specimen by analyzing the transmitted electrons that are elastically scattered by the sample. SEM gives predominantly surface information by analyzing reflected (back-scattered) electrons. STEM is a scanning technique that generates images by analyzing transmitted electrons.

In SEM, an electron beam is focused into a nanometer spot that is scanned over the sample. Interaction of the electrons with the sample leads to isotropic scattering in all directions (see Section 26.4). Back-scattered (reflected) electrons are detected to obtain topological information on the surface of the sample at resolutions between 0.1 mm to a few nanometer. SEM yields lower-resolution information than TEM (Figure 26.2), but offers the advantages of simple sample preparation and imaging the surface of large thick specimen. SEM is frequently used in material science, but small organisms and cells are also suitable subjects. Variants of SEM use charged particles other than electrons, for example helium ions.

TEM uses a wide electron beam to illuminate a µm-sized area of the sample at once, similar to the light beam in light microscopy. Electrons transmitted through the sample are then used either for imaging or for diffraction studies. The strong interaction of electrons with matter enables TEM to image the internal structure of tissues, organelles, molecules, and even atoms. The resolution of TEM ranges from 0.1 mm down to less than a nanometer, and increasingly rivals the resolution of crystallography and other high-resolution microscopy techniques (Figure 26.2). *Scanning transmission electron microscopy* (STEM) is a TEM technique in which a focused electron beam with a sub-nanometer diameter is scanned across the specimen, similar to SEM, but here the transmitted electrons are used to generate an image. STEM can also be used for spectroscopic applications, *e.g.* to analyze the elemental composition of the sample. In this case, STEM makes use of the ionization of atoms on the sample surface by the electron beam. The ejection of an electron leads to a subsequent intra-atomic electronic transition that generate X-rays with energies characteristic for the element that is being ionized (Figure 24.12, see Section 24.4.1), enabling its identification.

Sub-types of STEM include energy dispersive electron energy loss spectroscopy (EELS) that analyzes electrons for their energy loss due to inelastic scattering by the object, and energy-dispersive X-ray spectroscopy (EDX) in which the ionization of atoms by incident electrons produces X-rays.

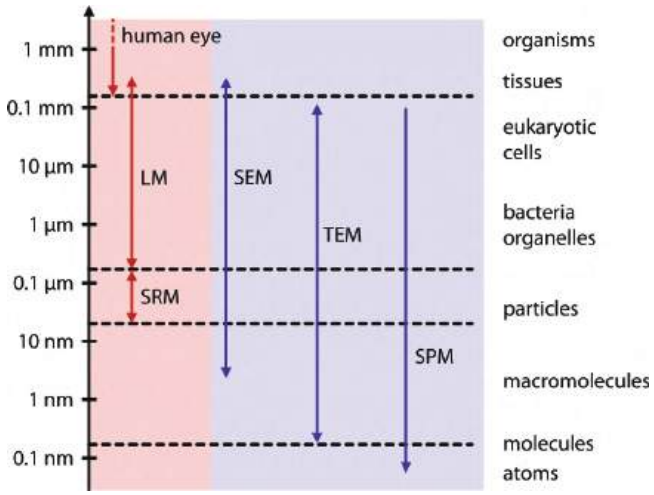


FIGURE 26.2: Resolution of EM compared to other microscopy techniques. The human eye can resolve objects down to 0.2 mm. LM: light microscopy, SRM: super-resolution microscopy (see Section 25.8), SEM: scanning electron microscopy, TEM: transmission electron microscopy, SPM: scanning probe microscopy (Chapter 27).

TEM images are inevitably two-dimensional projections of three-dimensional objects along the direction of the incident electron beam. When a series of two-dimensional projections is collected from the object in different orientations, the three-dimensional image of the object can be reconstructed from it. There are two major approaches for three-dimensional reconstruction that are based on two-dimensional image series obtained in different ways. In tomography, a thin slice of a sample is tilted with respect to the electron beam. In contrast, single-particle EM images an

array of randomly oriented particles deposited on a grid. We will discuss these two approaches in Section 26.5. Before we come to this, we will first familiarize ourselves with the main components of an electron microscope (Section 26.1). We will then treat the interactions of electrons with matter (Section 26.2), and the resulting requirements for sample preparation (Section 26.3). Section 26.4 then provides the physical basis for the formation of diffraction patterns and images. These Sections focus predominantly on TEM as it is suited for many biological applications.

26.1 OUTLINE OF AN ELECTRON MICROSCOPE

An electron microscope consists of three main parts: the illumination system, the imaging system, and the detection system (Figure 26.3). Electrons for illumination are produced by an electron gun, accelerated in an electric field, and focused onto the sample by a series of magnetic lenses (Section 26.1.3). More lenses, including objective, intermediate, and projective lenses, are used for image generation. While the focus of the projective lens system is fixed, the focal length of the intermediate lens can be adjusted, enabling switching between imaging and diffraction mode (Section 26.4). For detection, a fluorescent screen is used for visual inspection of images and diffraction patterns to adjust the instrument, whereas data collection uses an electron-sensitive detector. The entire electron path within the microscope (1–2 m), from electron gun to detector, is kept under ultra-high vacuum (10^{-5}–10^{-8} Pa; Figure 26.3). The evacuation of the microscope chamber serves to increase the mean free path of the electrons. At normal pressure, their mean free path is limited to a few millimeters due to frequent collisions with gas molecules, and they would never reach the sample. Under the ultra-high vacuum of the microscope, the mean free path of electrons is several kilometers.

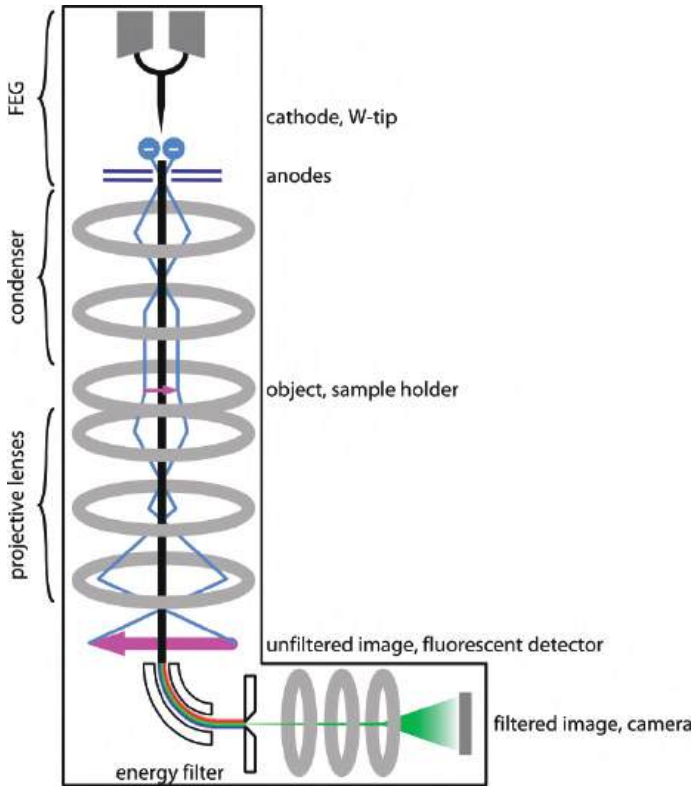


FIGURE 26.3: Schematic depiction of a transmission electron microscope. For illumination of the sample, an electron beam (blue) is generated by field emission from a cathode (black tungsten tip, W-tip). Electrons are accelerated in an electric field toward the anodes (dark blue). Beam shape, focus, and direction are manipulated by magnetic lenses (gray) and deflection coils (not shown). The condenser lens system focuses the beam onto the sample. For image generation, transmitted electrons are manipulated by a further lens system, and then detected by a detector in the image plane (large magenta arrow).

26.1.1 Generation of electron beams

Electrons for EM are generated by either thermionic emission or field emission. The aim is to generate a stable electron beam of high brightness and high coherence. Brightness of an electron beam is defined as emitted current per unit area per unit solid angle, measured in units of $A \cdot m^{-2} \cdot sr^{-1}$. Undesired fluctuations in the current reduce the stability of the beam. Coherence of the electron beam has two parts, spatial and temporal. A perfectly spatially coherent beam appears to come from a single point source and a perfectly temporally coherent beam has only electrons of a single wavelength and all are in phase. A temporally incoherent electron beam comprises a small spectrum of wavelengths, giving rise to an energy spread that is measured in eV.

Early microscopes used thermionic emission, where a filament made of a material with a low ionization energy (a low *work function*) serves as a cathode. Heating of the cathode to about 2000°C liberates electrons, very similar to the heated cathode in X-ray generation (Figure 24.12). Suitable materials include tungsten filaments and the mono-crystalline hexaborides of lanthanum (LaB_6) and cerium (CeB_6). Constant heating is associated with a loss of cathode material, which makes the electron beam unstable over time. The high operating temperatures also decrease the temporal coherence of the electrons: they have a range of wavelengths, or in other words an increased energy spread. The extended shape of the filament in addition leads to reduced spatial coherence. The three problems are substantially reduced by *field emission guns* (FEGs) that extract electrons at lower temperatures from a very sharp mono-crystalline tungsten tip of about 200 nm diameter by a strong electric field of several kilovolt between the cathode and the first anode in the microscope (Figure 26.3). FEGs come in two types, Schottky FEG and cold FEG. In the Schottky FEG, the tungsten tip is coated with a layer of zirconium oxide (ZrO_2), reducing the energy required to eject an electron from the surface by increasing the electrical conductivity of the tip. The combination of field emission and thermoionic emission at about 10^3 K in Schottky FEGs is more efficient at much lower temperatures compared to pure thermoionic emission. The latest development is the cold FEG, also known as a *cold cathode* because it is operated at room temperature. The tip is a needle-shaped tungsten crystal, which increases the efficiency of field emission. Cold FEGs produce beams with smaller diameters compared to (purely) thermionic electron sources (high spatial coherence) and smaller energy spread (ca. 0.3 eV), which is essential for high-resolution imaging. In addition, the brightness of the beam is up to 10^3 times higher compared to electron beams from thermionic emission. Electron beams from cold FEGs achieve a brightness of about 10^{12} $A \cdot m^{-2} \cdot sr^{-1}$.

26.1.2 Electron energy and resolution limit of EM

According to the Abbe limit, the highest resolution that can be achieved in microscopy is limited to half the wavelength of the incident light (Section 25.1.3). A simple view of the term resolution in microscopy is the smallest distance between two objects at which they can still be distinguished. Accordingly, the smaller the wavelength, the higher the resolution that can be obtained in principle.

Similar to photons, electrons can be described either as particles or as electromagnetic waves (wave-particle dualism, see Section 23.2.5.1). The wavelength λ of particles is given by the de Broglie relationship that we have encountered before in neutron scattering (eq. 23.43). The de Broglie wavelength depends on the impulse p, the product of the particle mass m and its velocity v, and the Planck constant h:

$$\lambda = \frac{h}{p} = \frac{h}{mv} \qquad \text{eq. 26.1}$$

For electrons of mass m_e traveling at a speed v_e, we obtain the wavelength of the electron λ_e as

$$\lambda_e = \frac{h}{m_e v_e} \qquad \text{eq. 26.2}$$

In the electron microscope, the electrons experience an acceleration by a voltage U_{acc}. As a result, their kinetic energy E_{kin} is

$$E_{kin} = U_{acc} \cdot e \qquad \text{eq. 26.3}$$

where e is the electron charge. In combination with the general definition of E_{kin}

$$E_{kin} = \frac{1}{2} m_e v_e^2 \qquad \text{eq. 26.4}$$

we obtain the wavelength λ_e of the electron as a function of the acceleration voltage:

$$\lambda_e = \frac{h}{\sqrt{2 \cdot e \cdot U_{acc} \cdot m_e}} \qquad \text{eq. 26.5}$$

In electron microscopy, electrons reach velocities approaching the speed of light c, making them relativistic particles. As a consequence, eq. 26.4 is no longer applicable to describe the energy of the electron. Instead, the relativistic impulse p_{rel} has to be considered, which is calculated from the *energy-momentum relationship:*

$$E_{rel}^2 = \left(p_{rel} \cdot c\right)^2 + E_0^2 \qquad \text{eq. 26.6}$$

Here, E_{rel} is the relativistic energy of the electron, $E_0 = m_e c^2$ is its rest energy, and m_e its mass at zero velocity. Solving eq. 26.6 for p_{rel} gives

$$p_{rel} = \frac{\sqrt{E_{rel}^2 - E_0^2}}{c} \qquad \text{eq. 26.7}$$

with $E_{rel} = E_0 + E_{kin}$, the sum of the rest energy E_0 and the kinetic energy E_{kin}, we obtain

$$p_{rel} = \frac{\sqrt{2E_0E_{kin} + E_{kin}^2}}{c} \qquad \text{eq. 26.8}$$

Substituting p_{rel} into eq. 26.1 then yields the wavelength of an electron traveling at relativistic velocity:

$$\lambda = \frac{h}{p_{rel}} = \frac{h \cdot c}{\sqrt{2E_0E_{kin} + E_{kin}^2}} \qquad \text{eq. 26.9}$$

Together with the kinetic energy expressed as a function of the acceleration voltage (eq. 26.3), we obtain the wavelength of a relativistic electron as

$$\lambda_e = \frac{h \cdot c}{\sqrt{2 \cdot m_e c^2 \cdot U_{acc} \cdot e + \left(U_{acc} \cdot e\right)^2}} \qquad \text{eq. 26.10}$$

For a comparatively low acceleration voltage of $U_{acc} = 10^5$ V (100 kV) we obtain a kinetic energy of $1.6 \cdot 10^{-19}$ J (equivalent to 100 keV). Such electrons correspond to electromagnetic waves with wavelengths $\ll$ 10 pm, several orders of magnitude smaller than that of light and about one order of magnitude smaller than chemical bond lengths. This is the reason why electron microscopy can, in principle, achieve atomic resolution (< 0.1 nm) whereas the resolution of light microscopy is limited to > 200 nm (eq. 25.16). The actual resolution achieved with electron microscopy techniques is, however, limited by the sample (Section 26.3), by the diffraction limits of the lenses, (Section 26.1.3; see Section 25.1.3), and by detector noise (Section 26.1.4).

26.1.3 Electromagnetic lenses

Electrons scattered by the sample can be viewed as electromagnetic waves, with different amplitudes and phases compared to the incident beam. By interfering with each other, they produce a scattering pattern, also termed a diffraction pattern, behind the object. We have encountered diffraction patterns from crystals by X-rays, neutrons, and electrons before (Sections 24.4, 24.7, and 24.8). From the amplitudes and the phases of a diffraction pattern the original image can be calculated by the help of the Fourier transform (Section 31.15). Here, "image" pertains to the object that was in the field of view of the beam that interacted with the sample. In optical microscopy, amplitude and phase information are combined into a magnified (inverted) image by lenses, which act as Fourier transformers. The corresponding elements in EM are *electromagnetic lenses*, or short *magnetic lenses*. Depending on how these lenses are operated, they can either produce an image or leave the diffraction pattern for further analysis (Section 26.4). Remember that in X-ray crystallography we cannot obtain an image directly: the phase information is lost because no lenses for X-rays exist, giving rise to the phase problem (see Section 24.1). A key advantage of EM compared to X-ray crystallography is the existence of magnetic lenses that allow us to work in image mode where the phase problem is absent.

The operating principle of magnetic lenses is based on the behavior of charged particles in an electromagnetic field: an electron passing an electromagnetic field experiences a Lorentz force perpendicular to its direction of travel, which alters its path. The Lorentz force is zero at the center of the lens, and highest at the periphery, leading to increasing deflection of electrons from center to periphery, and the focusing of all electrons in one spot at a characteristic focal length. In magnetic lenses, the magnetic field is generated by applying a current to a solenoid in the lens. Changing the current changes the strength of the magnetic field, which in turn changes the deflection of the electrons, and hence the focal length and magnification of the lens. Thereby, magnetic lenses enable the focusing and shaping of electron beams (Figure 26.4).

An electron microscope contains three larger arrays of lens systems: the condenser system, the objective lens system, and the projective lenses (Figure 26.3). The condenser system comprises 2–3 condenser lenses that shape the electron beam prior to its interaction with the object. A key element of the electron microscope is the objective lens system that comprises the sample holder where all electron-object interactions take place. The objective lens ensures homogeneous illumination of the object with a parallel electron beam in high-resolution TEM and focuses the electron beam into a spot for STEM (Figure 26.4).

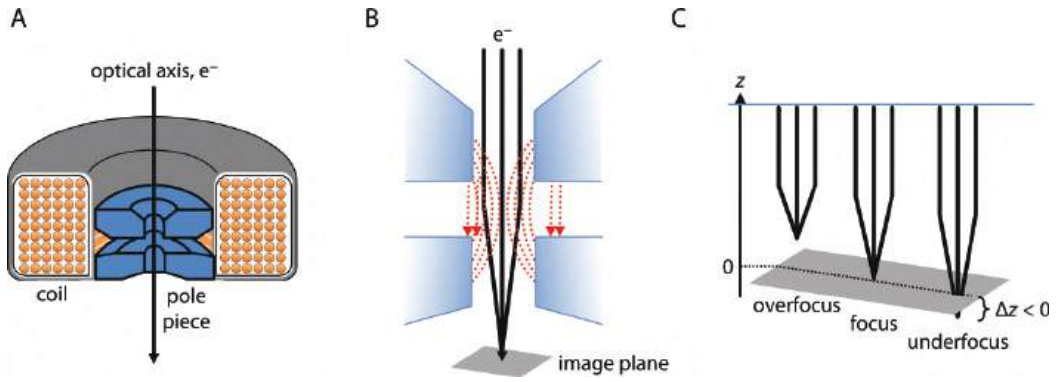


FIGURE 26.4: Electromagnetic objective lens. A: Cut-away of the copper wire coil (orange solenoid) encased in the soft iron housing (gray). The central pole piece (blue) has a bore to allow electron passage. B: The magnetic field between the two parts of the pole piece is shown as dashed red field lines. Electrons at the center are unaffected but those off-center are focused into a spot on the image plane (gray). C: Enlargement of the first intermediate image plane at the focus of the lens (middle). At small field strength and/or high electron velocity, the beam may be underfocused ($\Delta z < 0$; right) and the focal spot is below the image plane. With increasing field strength the lens reaches the focus ($\Delta z = 0$; in plane) and overfocus ($\Delta z > 0$).

The solenoid in objective lenses is a ring-shaped copper coil encased by a soft iron core. At the center of the coil resides the *pole piece*, a bipartite metal piece arrangement with a central hole or bore. An electric current through the copper wire establishes a magnetic field between the plates in the pole piece. Electrons traveling through the bore interact with the magnetic field, experience a Lorentz force, and are focused in a spot at the focal length of the objective lens. We will see later that EMs are often operated with the objective lens in underfocus or *defocus* mode (Sections 26.4.2, 26.4.3) to maximize image contrast. The image produced by the objective lens is then enlarged by the projective lenses and cast onto the detector.

The theoretical, diffraction-limited resolution r_{th} of a lens is a function of the wavelength λ and the half-opening angle θ of a cone of rays diffracted by a point object that can be imaged by a large, optically perfect lens (see also Section 25.1.3):

$$r_{th} = 1.22 \cdot \frac{\lambda}{\sin\theta} \approx 1.22 \cdot \frac{\lambda}{\theta} \qquad \text{eq. 26.11}$$

(see eq. 25.12). In electron microscopy, the scattering angle θ is small, ca. 0.01-0.05 mrad, and we can approximate $\sin\theta \approx \theta$. From eq. 26.11, we see that, the same way as for optical lenses (Section 25.1.1), the resolution of a magnetic lens is higher for smaller λ and larger θ. Magnetic lenses also have the same imperfections (aberrations) as optical lenses do, with the same consequences of image distortions and reduced resolution. Lens imperfections include *chromatic* and *spherical aberrations* (Figure 26.5), *astigmatism*, and axial *coma*. Chromatic aberration has its origin in low temporal coherence of the electrons. It is caused by the focusing of electrons with different velocities/energies into different focal planes. Spherical aberration is caused by different focal spots of electrons passing through the central and more peripheral regions of the lens, caused by inhomogeneities of the magnetic field over the diameter of the lens. Astigmatism arises when the lens focuses the beam more strongly in one direction, giving rise to ellipsoidal image distortions. Finally, axial coma arises when the illumination is not parallel to the optical axis of the objective lens, which leads to a tailing effect and image blurring. Deflection coils next to the electron gun and before and after the objective lens serve to correct some of these defects by tilting or shifting the beam such that it has only electrons of a certain energy and runs parallel to the optical axis. If lens imperfections are not corrected, the actual resolution achieved in EM remains below the theoretical limit.

The dominating non-ideality of magnetic lenses is the spherical aberration C_s, the dependence of the refraction of the beam by a lens on the distance from the center due to an inhomogeneous magnetic field. In a perfect magnetic lens, electrons passing through the center of the lens are unaffected by the magnetic field. The larger the distance of an electron from the optical axis, the stronger the refraction toward the axis by the lens. Spherical aberration causes additional bending of rays at the periphery of the lens, widening the focus of a point object to a disc with a diameter r_{sph}. The overall *practical resolution* r_{min} taking spherical aberration into account may be approximated by

$$r_{min} = r_{th} + r_{sph} = 1.22 \cdot \frac{\lambda}{\theta} + C_s \cdot \theta^3 \qquad \text{eq. 26.12}$$

where the spherical aberration coefficient C_s takes values of a few millimeters. Spherical aberration may be corrected for by a concave lens that spreads out the electron waves at the perimeter of the convex objective lens.

The second major type of non-ideality is the chromatic aberration C_c of a lens, which leads to different focal planes for electrons with different velocities/energies, increasing background noise and reducing the resolution of the images obtained. The energy spread of the initial electron beam is rather narrow, typically < 0.3 eV for cold FEGs, and it has a defined shape. The electron beam is temporally and spatially coherent. However, inelastic scattering from residual gas molecules in the microscope and, much more strongly, from the sample itself, deteriorates the beam. Inelastic scattering has two effects: it is associated with an energy loss of about ΔE = 15–25 eV, leading to an additional energy spread, and it leads to a further broadening of a point object into a disc of radius r_{chr}:

$$r_{\text{chr}} = C_{\text{c}} \cdot \frac{\Delta E}{E} \cdot \theta \qquad \text{eq. 26.13}$$

where C_{c} is the chromatic aberration coefficient. Chromatic aberrations due to inelastic scattering become severe as soon as the sample thickness exceeds just a few atom layers, which applies to virtually all biological specimens. The sample is thus the main factor limiting the resolution that can be achieved. As for spherical aberrations, the effects of chromatic aberrations can be corrected. Modern microscopes eliminate electrons that have lost energy due to inelastic scattering by the use of an *energy filter* (Figure 26.3). The energy filter is a magnetic prism, a type of electromagnetic lens curved by 90° that generates a magnetic field perpendicular to the flight path of the electrons. When passing, the electrons experience a Lorentz force that bends their trajectories as a function of their velocity (kinetic energy) onto a circular arc. An aperture behind the prism rejects inelastically scattered electrons and reduces C_{c} as much as possible before the image is recorded.

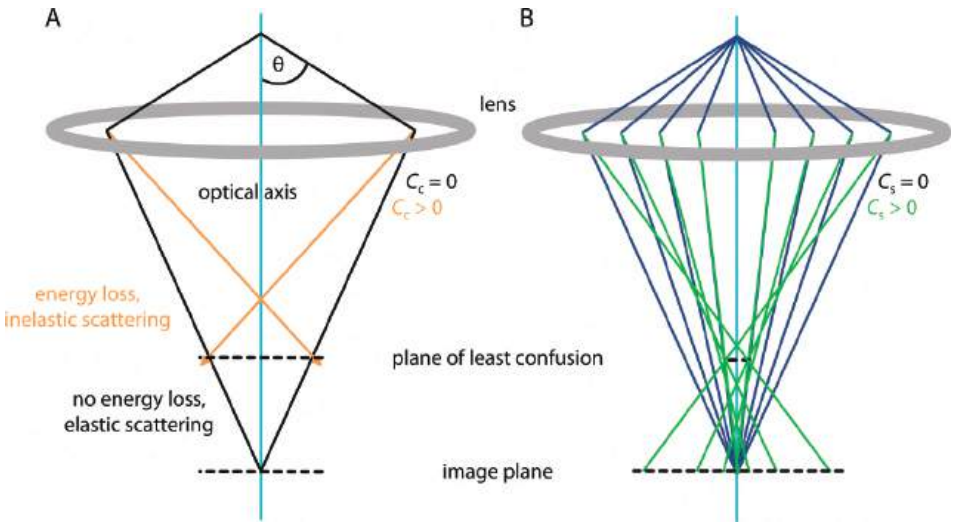


FIGURE 26.5: Chromatic and spherical aberration. A: Lenses with chromatic aberration focus electrons with a range of energies into different planes (dashed lines). Electrons that lost energy due to inelastic scattering by the sample (orange) are focused more strongly compared to elastically scattered electrons (black). B: In spherical aberration, the wave fronts are distorted such that their focal points are at different locations along the optical axis, each focal point defining its own focal plane. The smallest image of the object is a disc at the plane of least confusion, a compromise between all the other focal planes. The result of both, chromatic and spherical aberration, is imaging of a point object as a disc in the plane of least confusion (top dashed lines).

In certain applications, the energy loss is exploited to extract information about the sample. In *Electron Energy Loss Spectroscopy* (EELS), the separation of the electrons according to their energies by a prism is used to generate an intensity spectrum as a function of ΔE. The loss of energy during inelastic scattering depends on the electron density and hence the elemental composition, and thus the spectrum provides information on the electron distribution in the sample. EELS can achieve spatial resolution when combined with STEM. While EELS is frequently used in material science, it is not applicable to biomolecules because it needs much higher electron doses than these samples can tolerate.

26.1.4 Electron detection and detector noise

Electrons can be detected by photographic film, fluorescence screens, CCD cameras (Section 24.4.3), and *direct electron detectors*. Photographic film was the detector of choice for many years in TEM before CCDs became available. Fluorescence screens are mostly used to quickly generate images for visual inspection during instrument alignment. Data collection uses CCD cameras or direct electron detectors, a new class of detectors developed in 2010. Readout of CCD cameras is much faster than developing photographic films, but CCDs display poor lateral resolution. The current state of the art are direct electron detectors with increased detection efficiencies and reduced noise compared to CCDs (Figure 26.6).

The efficiency of signal recording in any detector is described by the *detective quantum efficiency* DQE, defined as the ratio of the squares of the output signal-to-noise ratio $(S/N)_{out}$ and the input signal-to-noise ratio $(S/N)_{in}$:

$$DQE = \frac{\left(\frac{S}{N}\right)^2_{out}}{\left(\frac{S}{N}\right)^2_{in}} \qquad \text{eq. 26.14}$$

A perfect electron detector would have a DQE of unity over its entire surface and independent of the energy of the electron detected. Such a detector would not add additional (electronic) noise to the signal, and it would count each and every single electron with equal weight. However, this is not achievable with current techniques, and DQE thus is always < 1, and is a function of the energy of the electrons detected.

Photographic film and CCD cameras have approximate DQEs of 0.3 and 0.1. The relatively large DQE and the small pixel size of photographic film explain why historically the time-consuming film development was still used in TEM when optical microscopy techniques had already adopted CCDs. Direct electron detectors reach DQEs of about 0.5. An increase of DQE to 0.5 might not sound impressive, but the associated increase in resolution achieved by direct electron detectors was paramount for the "resolution revolution" in EM (Section 26.5).

Direct electron detectors consist of an array of semi-conducting pixels, with each individual pixel coupled to its own readout electronics that enables data readout from all pixels in microseconds. Energy-sensitive detectors (*e.g.* the Falcon) integrate the energy deposited, while electron-counting detectors (*e.g.* the K2) count the number of electrons that hit the detector, irrespective of their energy. The pixel array is arranged on a support matrix, which has the unfortunate property to back-scatter electrons towards the pixel array, introducing electronic noise. To minimize this noise, the support matrix is kept as thin as possible by a manufacturing process called *back thinning*, setting the overall thickness of the detector to only about 50 µm.

The pixels of direct electron detectors are quickly saturated at electron fluxes exceeding 1–10 e s^{-1}. To circumvent saturation effects, all direct electron detectors operate in *movie mode* to spread the energy/number of the recorded electrons over several hundred images per second. The high DQE and frame rates (images per second) of direct electron detectors have the added benefit that, compared to CCD detectors, a lower electron dose is sufficient to record complete data, reducing radiation damage (Section 26.2.3).

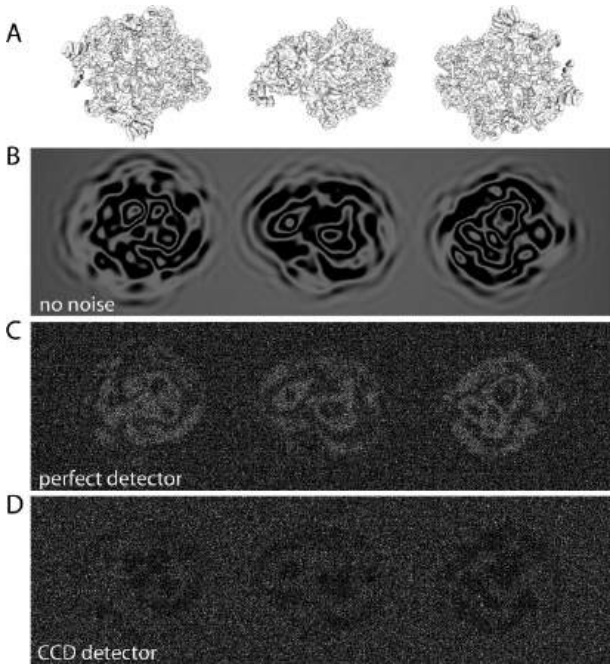


FIGURE 26.6: Detection of EM images. The structure of the mitochondrial ribosome ("mitoribosome"; PDB-ID 5aj4), determined at a resolution of 0.38 nm, was used for these projection calculations. A: Three orientations of the ribosome. B: The projection of the orientations in the absence of any noise. C: EM image from a perfect detector with DQE = 1. D: Image detected by a CCD detector. The projections of the molecules are often barely visible by eye in the images.

The image on a pixel array detector is a digitized version of a continuous object, essentially a table of intensities per pixel. The maximum resolution of such a digitized image is limited by the image pixel size, which depends on the (fixed) detector pixel size and the (chosen) magnification of the microscope. The image pixel size is therefore a parameter value set before data collection. If a large magnification is chosen, the image pixel size is small with a concomitant small field of view, and *vice versa*. The physical resolution that the detector is capable to provide by setting the magnification has nothing to do with the actual resolution obtainable from a sample. An optimal magnification setting is a compromise between a maximum number of particles in the field of view and the sample-limited resolution.

In EM, resolutions are often stated as inverse distances or frequencies q. The *Nyquist frequency* q_{Nyquist} is a function of the image pixel size:

$$q_{\text{Nyquist}} = \frac{1}{2 \cdot \textit{image pixel size}} \qquad \text{eq. 26.15}$$

The relation in eq. 26.15 is also known as the *Nyquist–Shannon sampling theorem*, which states that the inverse of the image pixel size, also called the sample rate, is twice the sampling frequency of data acquisition (q_{Nyquist}). In other words, the resolution of the image cannot exceed twice the pixel size. For example, if we use a magnification that images an object onto a detector such that each 0.2 nm (2 Å) of the object spans a single pixel of the detector, the image pixel size is 0.2 nm. According to eq. 26.15, the Nyquist frequency is then 2.5 nm^{-1} (0.25 Å^{-1}). The inverse of the Nyquist frequency, 0.4 nm (4.0 Å), is the maximum achievable resolution of the image (and the final reconstruction), called the *Nyquist limit* or *Shannon sampling limit*. Imaging the same object onto the same detector, but with a two-fold lower magnification, gives an image pixel size of 0.4 nm, a Nyquist frequency of 1.25 nm^{-1} (0.125 Å^{-1}), and a maximum resolution of 0.8 nm (8.0 Å). To compare different detectors with different pixel sizes, the frequencies q are often reported as multiples of the Nyquist frequency.

26.2 ELECTRON-MATTER INTERACTIONS

Before we discuss the preparation of samples for EM (Section 26.3), we first need to understand how electrons interact with the sample. When electrons interact with objects without being absorbed, they may be scattered either elastically or inelastically, leading to a change in traveling direction. In elastic scattering, the energy of the electrons does not change, whereas inelastic scattering is associated with a change in their kinetic energy. In EM, elastic scattering is the process that enables image formation (Section 26.2.1); electrons scattered inelastically do not contain structural information. Electrons can also cause damage to the sample by ejecting other (secondary) electrons from an atom, thereby producing X-rays (Figure 24.12), visible light, or Auger electrons (Section 26.2.2, Figure 26.1), by generating radicals that initiate detrimental chemical reactions, or by physically moving the object under study. These effects, collectively referred to as radiation damage, lead to blurring of the images obtained and a concomitant loss of high-resolution information during the EM experiment (Section 26.2.3).

26.2.1 Elastic scattering of electrons

Elastic scattering is the desired process in TEM: the electrons scattered elastically by the sample contain structural information. Elastic scattering occurs when the electrons are deflected by the nuclei of the object, and change their direction of travel without losing energy. The deflection depends on the Coulomb force that the electrons experience, and on the time they are affected by this force. The Coulomb force is higher the closer the electron passes to the nucleus and the larger the charge (atomic number) Z of the nucleus. The time this force is experienced is larger for slower than for faster electrons. Scattering angles θ are usually in the range of 0.5–5° (ca. 8.5–85 mrad). The scattered electrons, when viewed as an electromagnetic wave, are phase-shifted relative to the incident electrons. We will see later in this Chapter where this phase shift originates (Section 26.4.1) and how it relates to image formation (Section 26.4.2).

The scattering power of an object is quantified by the electron scattering factors of its composing atoms. The approximate scattering power f_e of an atom for electrons may be calculated from its atomic scattering factor f (see Section 23.2.1) by the Mott-Bethe formula

$$f_e(\theta,Z) \approx \frac{\lambda^2}{128\pi^4 a_0} \cdot \left(\frac{Z - f(\theta,Z)}{\sin^2\theta} \right) \qquad \text{eq. 26.16}$$

where Z is the atomic number of the electron-scattering element, λ is the wavelength (energy) of the electron, and a_0 is the Bohr radius of 53 pm, the most probable distance between the proton and the electron in a hydrogen atom at ground state. Remember that the X-ray atomic form factor is the Fourier transform (Section 31.15) of the electron density of the atom (Section 24.4.2; Figure 23.7) and the experimental result of an X-ray crystallographic experiment is the electron density distribution. Quite analogously, the electron scattering factor of an atom is the Fourier transform of the Coulomb potential density, and the result of an EM experiment is the Coulomb potential density distribution. The Coulomb potential of an atom is dominated by the nucleus, with some contribution by the electron cloud. In Section 23.2.1 we defined the molecular form factor for X-ray scattering in ideal solutions as the sum of all atom contributions. An analogous molecular form factor for the non-ideal case of solids, *i.e.* crystals (Section 24.4.2) or EM samples, cannot be calculated.

We see from eq. 26.16 that, at a given scattering angle θ, the scattering power of an atom is proportional to the atomic number Z. Coating the sample with atoms of high atomic numbers thus boosts elastic scattering and the signal from the sample. This is exploited in metal shadowing (Section 26.3.2) and negative staining (Section 26.3.3) of EM samples using heavy atoms. Unfortunately, biological samples are predominantly composed of the lighter elements H, C, O, N, P, and S, meaning that elastic scattering from biological samples is small and hence difficult to measure.

26.2.2 Inelastic scattering of electrons

Inelastic scattering is associated with an energy loss of the electrons. The energy difference is transferred onto the sample, and the stronger the inelastic interaction, the more energy is transferred. The energy transfer from the incident electrons onto the specimen inevitably damages the sample, leading to physical movement of atoms and chemical reactions (Section 26.2.3). Inelastic scattering can also lead to the production of heat, visible light, secondary electrons, and X-rays (Figure 26.1). All these processes lead to blurring of the recorded images. High-resolution image generation by TEM therefore seeks to minimize inelastic scattering (Figure 26.7).

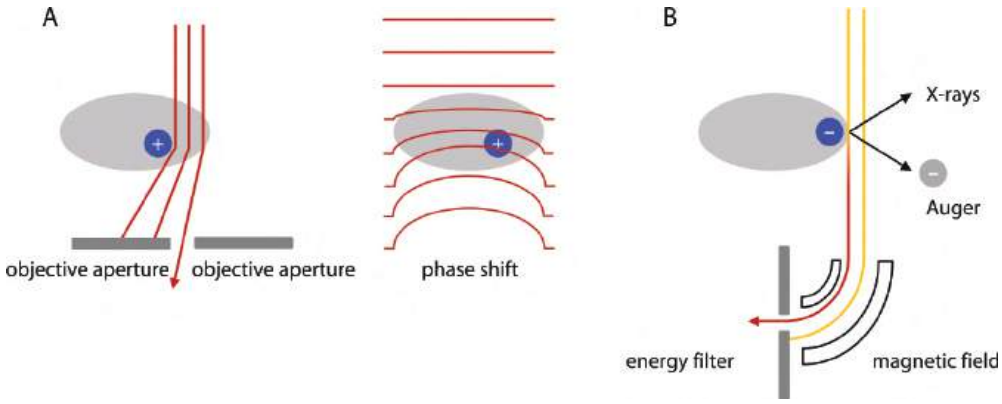


FIGURE 26.7: Elastic *versus* inelastic scattering. A: Elastic scattering of electrons by nuclei. The electrostatic interaction causes a deflection of the electrons without loss of energy (left). In TEM, an objective aperture (dark gray) rejects the most highly deflected rays, as these often come from inelastic scattering. The slightly deflected rays are phase-shifted relative to the incident beam and interfere with the unobstructed beam (right). B: Inelastic scattering of electrons by surface electrons leads to a reduction in electron energy, velocity, and wavelength (indicated by the color change from yellow to red). In a magnetic field, the slower inelastically scattered electrons are deflected more strongly, and an aperture (the energy filter or velocity selector) selects for the wavelength of the elastically scattered electrons that are detected. Upon inelastic scattering of electrons by surface electrons, X-rays can be emitted, or secondary and Auger electrons can be generated (see also Figure 26.1).

The probability of inelastic scattering increases with higher atomic numbers of the scattering elements, thicker samples, and smaller acceleration voltages, *i.e.* smaller kinetic energy of the incident electrons. Preparation of thin samples and increased acceleration voltage are thus two possible measures to minimize inelastic scattering. Remaining inelastically scattered electrons have less energy than elastically scattered electrons and are removed by two methods, an objective aperture after the sample and an energy filter (also known as velocity sensors) placed in front of the detector.

26.2.3 Radiation damage

The collective term for changes to the sample by the electron beam is radiation damage. Radiation damage is the main source of decreased signal in EM. Unfortunately, it is impossible to avoid radiation damage in high-resolution TEM. Some of its effects, however, can be removed by computational image analysis and processing.

A main form of radiation damage is the movement of areas of the sample. Analysis of the individual electron micrographs (frames) in EM movies showed early on that specimens exposed to an electron beam move in patches of 0.3–0.5 µm and may rotate by a few degrees. This movement occurs on a typical time scale of 1 $nm \cdot s^{-1}$, which is large on the time scale of image recording. Specimen movement is present even in cryogenic samples where the objects under study are embedded in vitreous ice (Section 26.3.4).

Beam-induced object movement is thought to be caused by a build-up of positive charge by ejection of electrons from the sample. The positive charges change the interaction of the electron

beam with the sample. The build-up of charges can be reduced by suspending the sample on sheets of electrically conductive graphene. Object movement during data acquisition can be corrected for by computational methods if images are recorded in movie mode. From the image stacks, the position of an object can be followed across 10^4–10^5 frames, defining its trajectory. With this trajectory, the position of this object is then back-calculated to the original position in each frame of the movie. The corrected frames are summed to the final image, leading to an increased signal-to-noise ratio. The procedure is improved further when several objects are in the field of view of the microscope, so that several trajectories can be determined simultaneously and used for back-calculation. Deep neural networks (Section 20.3.2) trained with previously collected EM data of entirely different samples showing object movement have also proved useful for back-calculation of new micrographs (Section 26.5.1). The development of algorithms for back-calculation of shifted objects in movies was the second major improvement that contributed to the "resolution revolution" in EM.

A second major class of radiation damage manifests as chemical damage to the sample. Inelastically scattered electrons produce radicals from water and organic components, much in the same way as X-rays do (Box 24.2). Hydroxyl radicals and nascent hydrogen in particular are very reactive, chemically degrading the object. The generation of radicals by electrons is not temperature-dependent, but the chemical reactions they initiate are. To reduce the temperature-dependent effects of radiation damage, samples are cryo-cooled and vitrified into a glass-like state (cryo-EM; Section 26.3.4), similar to cryo-protection in X-ray crystallography (Box 24.2).

Electron-induced damage can also be minimized by optimization of parameters for image acquisition. It is common practice to establish optimal microscope parameters at uninteresting regions of the sample where damage is unimportant, reserving the interesting regions for fast data collection. One possibility is the reduction of the overall electron dose of the incident beam, measured in electrons per·nm^2 in EM (instead of Gray (J·kg^{-1})). Significant radiation damage sets in at about 2500 e·nm^{-2}; at doses exceeding 50000 e·nm^{-2} the sample disintegrates. While low electron doses minimize radiation damage, they also reduce the signal-to-noise ratio and thus the resolution achievable. The electron doses used for the measurement are therefore a compromise: in high-resolution single particle EM, doses are limited to ca. 2000 e·nm^{-2} (20 e·Å^{-2}) to minimize damage and achieve high resolution. Electron tomography often uses higher doses of ca. 8000 e·nm^{-2} where some radiation damage occurs and limits the resolution attainable, but the effect is (partially) compensated by the increased signal-to-noise ratio at this dose and often the lower resolutions obtained in tomography are sufficient to answer the biological question.

A further parameter that can be adjusted is the acceleration voltage U_{acc}. A higher acceleration voltage leads to high-energy electrons that show less inelastic scattering than lower-energy electrons. Unfortunately, with increasing U_{acc} elastic scattering is also reduced because the interaction time of the faster electrons with the atoms is shorter, which leads to an overall reduced signal and contrast. Again, a compromise is needed. In modern cryo-EM, the sweet spot of U_{acc} = 300 kV is often used.

26.3 SAMPLE PREPARATION FOR EM

Preparation of samples for EM serves three main purposes: increasing the interaction of electrons with the sample to improve contrast, stabilizing the sample in the vacuum inside the microscope chamber, and minimizing inelastic scattering by the sample that leads to blurring of EM images. Improved interaction of specimens with electrons is achieved by metal shadowing and negative staining (Sections 26.3.2 and 26.3.3). To minimize dehydration and structural damage to biological samples in the vacuum of the EM, they are immobilized in amorphous ice (Section 26.3.4), embedded in resins, or fixed by chemical crosslinking. Inelastic scattering is reduced by preparing thin samples. Sample preparation is no easy task, because the requirements are in conflict with each other. For example, a thick sample is beneficial for image contrast, but also enables stronger interactions of the electrons with the sample, leading not only to more elastic scattering, but also to increased inelastic scattering and radiation damage.

26.3.1 EM GRIDS AND SAMPLE SUPPORTS

The "sample holder" for EM consists of a metal grid, covered by a sample support. The metal grid, made of Cu or Au, has a diameter of about 3 mm, and a mesh size of 30–100 µm (Figure 26.8). The sample support can be a metal coating of Au or Pd/Ni that generates a regular array of holes or *holey grid* on the mesh with hole sizes between 1.2 µm and 5 µm. Alternatively, the sample support may be a 5–10 nm thick layer of graphite or graphene (carbon film) with randomly dispersed holes or no holes at all. The advantage of holes is that electrons may reach objects embedded in the holes without passing through the grid material, which would unnecessarily absorb and scatter electrons to increase background noise. Placing a droplet of a buffered sample on the sample support allows distribution of the molecules into the holes. The size of the holes is chosen according to the magnification desired. Ideally the field of view of the microscope covers a single hole. High magnifications result in a smaller field of view of the microscope, and require smaller hole sizes. For example, 1.2 µm and 5 µm holes are suitable for magnifications between > 50000 and 20000.

The nature of the sample support determines the interaction of the sample with the grid. Grids with a carbon film are quite hydrophobic so that *e.g.* membrane proteins adhere non-specifically at low concentrations. For the application of more hydrophilic samples, the grid is made hydrophilic in a partially evacuated chamber that contains residual air or O_2/Ar mixtures. When a high voltage is applied, ions are formed by electrical discharges. The plasma formed transfers charges onto the grid and simultaneously destroys organic impurities by *plasma cleaning*. The grids thus prepared are ready to accept a sample within a few minutes or they lose their charge again by attracting ions and impurities from the air. By adding other volatile substances during the electric discharge, the surface properties of the grid can be modified further, for example to prevent deposition of particles in specific orientations. This becomes important for single-particle cryo-EM (Section 26.5.2), where a well-prepared sample consists of a grid with a large number of separated, evenly distributed particles in all possible orientations, so that as many different projections of the particles as possible can be imaged from a single grid.

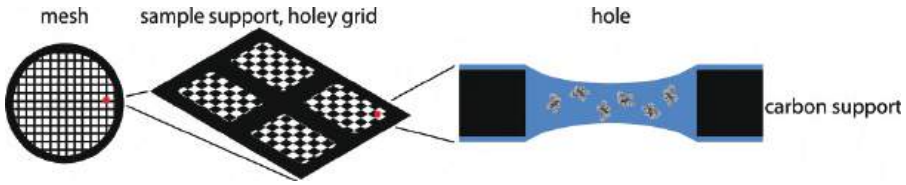


FIGURE 26.8: Sample supports for cryo-EM. The grid of 3 mm diameter on the left contains 40 µm holes, four of which are enlarged in the center showing a regular array of holes or holey grid. The right-hand side panel shows one of these small holes filled with a protein sample under cryo-conditions. Vitreous ice (blue) of up to 100 nm thickness keeps the objects in position.

26.3.2 METAL SHADOWING FOR SEM

SEM images the surface properties of an object. To enhance surface features of a sample and to increase the contrast in the electron micrograph, samples for SEM are coated with metals, and the resulting replica is imaged. To preserve their three-dimensional structure, biological macromolecules are stabilized by addition of 40–70% glycerol before metal shadowing. Shadowing uses metals of high density that strongly scatter electrons at minimal deposit thickness. The metal coat must be finely granular, uniform, and amorphous in structure and be chemically and physically inert under the experimental conditions. Metals fulfilling these requirements are Pt, W, Ta, and Eu. The metal vapor is evaporated from a tungsten filament and deposited on the sample surface at an oblique angle of a few degrees. The small angle aims to predominantly deposit metal onto the high points of the sample in order to highlight these contours in the replica image. This means that for smaller objects the angle of shadowing must be more oblique. Unfortunately, the smaller the angle, the more are depressions on the surface filled from one direction. Surface features of the object are distorted or removed entirely, leading to replica that do not represent the original object. Artifacts

from metal shadowing can be severe and include vanishing of handedness in nucleic acids and artificial "holes" in highly symmetric oligomers. For a three-dimensional rendering of the specimen, the sample may be rotated during shadowing. Areas where the metal coating is thickest will scatter electrons strongly, creating dark areas in the image, while areas with thin metal coating are transparent to the electron beam and appear bright in the image. If the coating is thin, imaging may also be possible in transmission mode.

26.3.3 Negative staining for TEM

To enhance the contrast of TEM images, proteins, protein complexes, and biological membranes are stained with heavy atom salts, most commonly uranyl formate or uranyl acetate. In contrast to metal shadowing, the heavy atom salt in negative staining also penetrates the sample, and its interior structure can be imaged. The staining needs to be done carefully to avoid artifacts such as sample desiccation that leads to object distortion. The negatively stained samples are stable for weeks at room temperature. Again, the contrast medium is imaged, not the original object. The staining leads to strong electron scattering such that the "objects" appears bright on a dark background at high contrast. Too much contrast medium, however, not only enhances contrast, but increasingly fills depressions on the surface of the object, and these surface features are lost in the image. Too little contrast agent, on the other hand, will not lead to maximized contrast, and will leave the object vulnerable to radiation damage (Section 26.2.3).

Negative staining is useful for detecting low-resolution features of about 2 nm and to quickly judge the purity, oligomeric state, conformational homogeneity, and orientation distributions of single particles on a grid for later TEM experiments with unstained samples.

26.3.4 Native samples in vitreous ice

High-resolution EM studies require unaltered samples that must be at least somewhat resistant to the high vacuum and to radiation damage. The basic principle is to form a thin layer of the aqueous sample on a grid and then to vitrify the water by rapidly plunging the grid into a cryogenic liquid (cryo-cooling). Vitrification avoids crystallization of water, which would not only reduce the signal by changing the native structure of macromolecules, but also increase the noise by strongly scattering the incident electrons. The high cooling rates required to achieve water vitrification in the sample are provided by cryogens of high heat capacity and thermal conductivity such as liquid propane, ethane, or mixtures of the two. Depending on the size of the sample, these liquids allow for cooling rates of 10^3–10^4 K·s^{-1} and they also minimize the generation of thermally insulating gas bubbles (*Leidenfrost effect*) around the sample during vitrification, which decrease the cooling rate. The sample itself must be less than a few hundred nanometer thick or the cooling rates will decrease too much and ice will form. Thin layers are formed by placing a droplet containing the object on a holey grid, waiting until the objects have settled into the holes, and removing surplus liquid by blotting with filter paper. Alternatively, self-wicking grids are available, or nanometer-sized sample droplets are "printed" directly onto the grid without the need for blotting. Larger sample volumes may sometimes be vitrified after addition of cryo-protectants that slow down ice crystal formation. Popular cryo-protectants are poly-alcohols such as dextrose, dextrin, dextran, trehalose, glucose, or glycerol. However, these chemicals inevitably increase the density of the amorphous solvent and thus reduce the contrast between solvent and object. For example, the difference in densities for proteins and amorphous water is already small (1.4 g·cm^{-3} *versus* 0.94 g·cm^{-3}), so the contrast between sample and water is also weak. Cryo-protectants worsen this trend. To avoid cryo-protectants, larger objects of 10–300 μm diameter, including viruses, single cells, intracellular structures, and tissues, can be vitrified by high-pressure vitrification (often called "high-pressure freezing"). The cooling rates under normal pressure are too small to avoid ice crystal formation at the center of thick samples. At pressures above 0.2 GPa, ice crystals form much more slowly than under ambient pressure, allowing cryo-protection of thicker samples without additional cryo-protectants.

Whole cells and tissue sections with a thickness of several µm are intransparent for electrons and thus need thinning after they have been vitrified. Thinning is achieved by abrasion of the vitrified sample, until a *lamella* of ≤ 500 nm thickness remains (Figure 26.9). In focused ion beam (FIB) milling, a 5–10 nm diameter beam of ions in a vacuum bombards a small part of the cryo-cooled sample grid and removes layer by layer of material both from the top and the bottom (after flipping of the grid). FIB milling is an integral part of SEM, where the thinning process can be monitored directly within the microscope. For TEM samples, it is performed in a separate instrument. Afterwards, the resulting lamella is imaged from different angles, and images are combined into a tomogram (Section 26.5.1).

Liquid Gallium is most widely used as the ion source (Ga^+) for the FIB because of its low melting point, low volatility, and low vapor pressure, properties that allow convenient production of low-energy Ga^+ ions by field emission. However, Ga^+ ions can react with samples and alter their structure, therefore ions produced from gases (N_2, O_2, Ar, Xe) have been used as well. The ionic radius, velocity, and intensity of the ion beam determine the rate of abrasion. The milling process can be started by abrasing thicker layers quickly with higher-energy ions and wider ion beams until the biologically interesting layers are reached. At this point, the ion beam intensity, its energy and/or its diameter are reduced for finer abrasion of thinner layers to fine-tune lamella preparation. To minimize damage to biological samples by the ion beam, a thin protective layer of organometallic platinum may be deposited on the sample by a gas injection system prior to milling. The gas injection system, located next to the FIB, deposits the metal layer by way of the temperature gradient between gas (ambient) and sample (cryogenic). The next round of FIB milling then erodes the metal layer and the top layer of the sample, enabling finer/gentle sample abrasion.

The location of the interesting layer of the sample is unknown *a priori*. As a consequence, the milling process needs to be monitored closely, which makes it time-consuming. Fluorescent labels in the sample can act as *fiducial markers* to label regions of interest such as membranes, membrane proteins, organelles, or sub-cellular structures for which a biological marker exists. Fluorescence microscopy can then help select promising areas for the milling process.

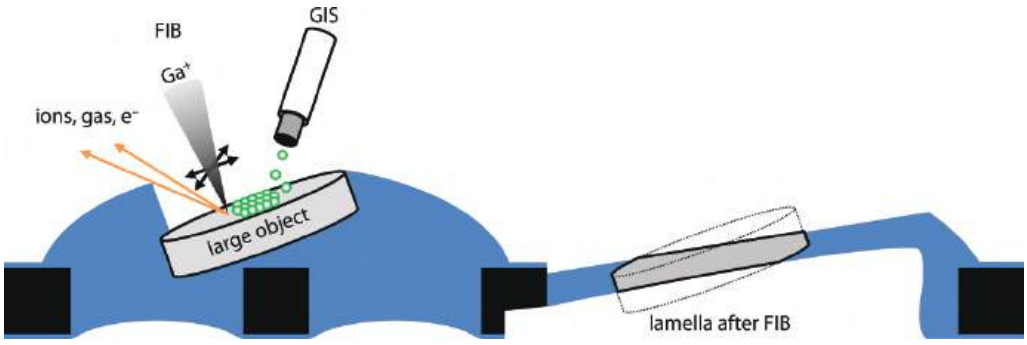


FIGURE 26.9: FIB milling for cryo-EM. Schematic representation of a larger, electron intransparent object (gray) in vitreous ice (blue) that needs milling into an electron-transparent lamella with a focused ion beam (FIB) for TEM sample preparation. The gas injection system (GIS) next to the FIB deposits a protective platinum coating onto the vitrified sample to reduce excessive erosion by the FIB. A FIB of a few nanometers diameter using Ga^+ ions in this example is then scanned across the sample and evaporates both the protective coat and a thin layer of sample matter as gas, ions, or secondary electrons.

26.4 TEM IMAGING AND DIFFRACTION

Transmission electron microscopes operate in either imaging or diffraction mode. Just the same way as in crystallography, where the electron density represents the object in real space and the diffraction pattern corresponds to reciprocal space, the imaging mode in EM collects real space data, whereas the diffraction mode collects data in reciprocal space. Correspondingly, the diffraction mode is used for analyzing ordered samples such as fibers and crystals (see Section 24.8), whereas the imaging mode is used for most other biological applications. Both diffraction pattern and image

are present simultaneously in a TEM. Which of the two is recorded on the detector essentially depends on the focal length (or strength) of the intermediate lens and the position of apertures in the electron path of the microscope (Figure 26.10).

In imaging mode, the objective lens together with an objective aperture placed at or near the *back focal plane* (Figure 25.1) produces a first intermediate image. This image is formed on the distant side of the objective lens just in front of the intermediate lens. The focal length of the intermediate lens is set to a small value (it is made strong) to magnify the object in a second intermediate image. Finally, a set of projective lenses further magnifies the second intermediate image in several steps until the final image is recorded on the detector.

For diffraction mode, the objective aperture is removed from the back focal plane of the objective lens. Instead, a selected area aperture is placed at the focal plane of the first intermediate image. This aperture limits the area from which the diffraction pattern is collected; the approach is therefore termed *selected area electron diffraction* (SAED). The focal length of the intermediate lens is increased (its strength is decreased) to focus the diffraction pattern into the plane of the second intermediate image (the back focal plane of the intermediate lens). The diffraction pattern is then magnified by the projective lens system and cast onto the detector.

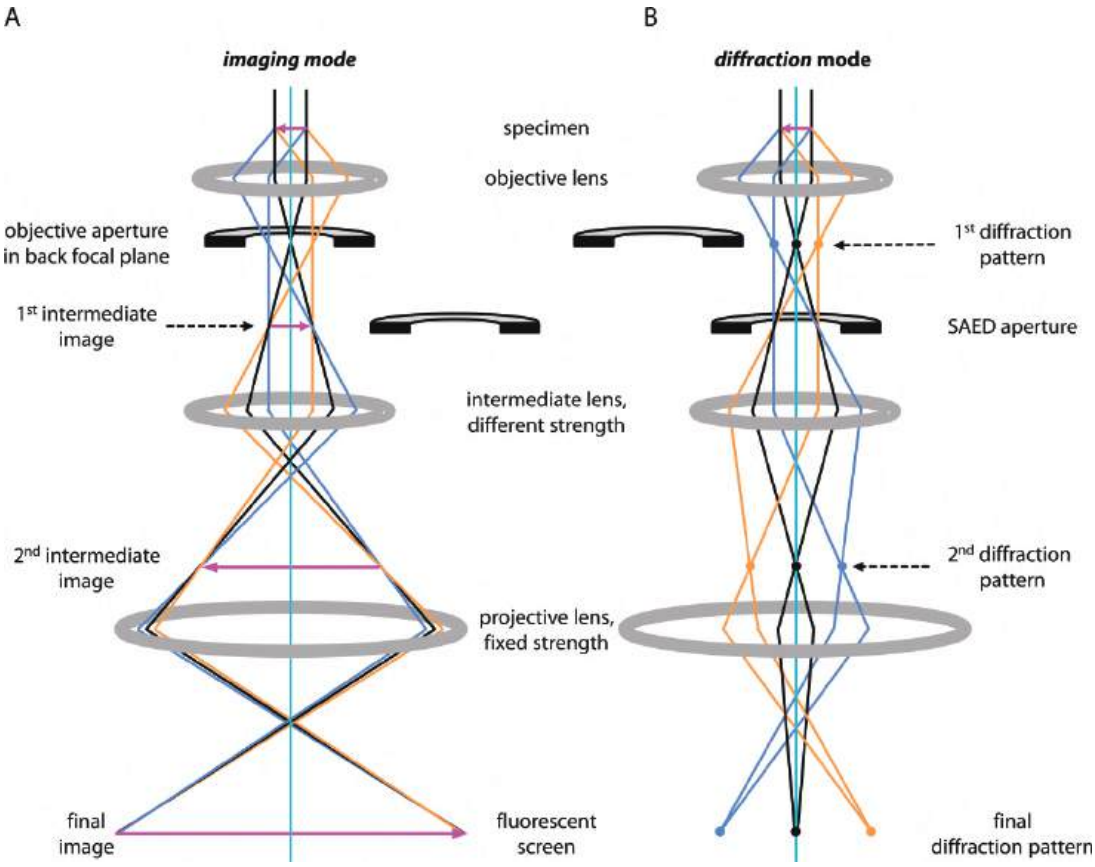


FIGURE 26.10: Modes of operation in TEM. A: In imaging mode, an object (magenta arrow) is magnified *via* various intermediate images onto the detector. The objective aperture is in the place of the back focal plane where the first diffraction pattern forms and the intermediate lens is set to short focal length. B: In diffraction mode, the objective aperture is removed from the back focal plane of the objective lens, and a selected area electron diffraction (SAED) aperture is inserted at the position of the first intermediate image. The intermediate lens is set to long focal length to magnify the diffraction pattern (colored circles) in the plane of the second intermediate image. In both cases the final image is magnified on the detector by the projective lens system. In contrast to the intermediate lens, projective lenses have fixed focal lengths.

The key elements that enable formation of the image or diffraction pattern are the phase shift the electrons experience on scattering, and the interference of scattered waves with the incident beam. We will take a closer look at how electron scattering by an object produces diffraction patterns or images in Sections 26.4.1–26.4.3.

26.4.1 Plane waves, defocus, and phase shifts

The generation of images and diffraction patterns can be explained by viewing the object of interest as an array of atoms that act as individual scattering centers. Each of these scattering centers interacts with the incoming electron wave and generates a scattered wave, the *wave front* or *component wave*. These scattered waves are distorted: they have a phase shift and a change in amplitude relative to the incoming beam. The wave fronts from all scattering centers then interfere with each other and with the incoming beam. At certain scattering angles θ positive interference occurs, leading to measurable scattering intensities (Figure 26.11). The position and intensities of the scattered beams is a function of the defocus value Δz, the distance of the focal point of the incoming electron beam from the first intermediate image plane (Figure 26.4). At certain defocus values Δz the extent of interference between the scattered wave fronts is maximal, hence the change in amplitude is maximal as well. As we will see, Δz is an important setting during TEM data collection (Section 26.4.3) because in the end only amplitudes can be measured.

The result of the interference is a diffraction pattern that is in focus in the back focal plane of the objective lens. This diffraction pattern is the Fourier transform (FT) of the object. As usual for a FT, its units are reciprocal to the units of the object it was calculated from. If we assign the unit nm to the distances between scattering centers in the object, the Fourier transform of the object has *reciprocal distances* of units nm^{-1}. These distances are called *spatial frequencies* q:

$$q = \frac{4\pi}{\lambda}\sin\theta \qquad \text{eq. 23.30}$$

We have introduced the concept of spatial frequencies and their relation to the scattering angle θ before in small angle X-ray scattering (Section 23.2.2). According to eq. 23.30, $q = 0$ corresponds to the direction of the incoming beam ($\theta = 0$), and large values for q correspond to large scattering angles. High-resolution information is stored in beams diffracted at high angles θ and is visible in the diffraction pattern as intensity at high spatial frequencies q. Waves scattered at large angles always contain the high-resolution information. This is the same in X-ray crystallography (Section 24.4.2) where, somewhat confusingly, the resolution is stated as distance, not as inverse distance ("frequency" in EM).

The diffraction pattern in the back focal plane contains amplitude and phase information, separated as the real and imaginary parts in the Fourier transform of the object (see Section 31.11). However, electron detectors are not phase-sensitive, and therefore the diffraction pattern recorded by the detector purely reflects the amplitudes. Similar considerations apply to EM imaging: the information about an object that is contained in scattered electrons is mainly encoded by phase differences, not amplitude differences. An important consideration in EM is therefore to "transfer" the unmeasurable phase information into measurable amplitude information.

To summarize, the phase shifting by the elastic scattering poses two challenges in EM: (1) maximizing the information encoded in the phases (phase contrast) by optimal instrument settings, and (2) converting the phase information as best as possible into recordable amplitude contrast (Section 26.4.3). We will discuss next why phase contrast is essential in EM and how it may be maximized and converted into amplitude contrast.

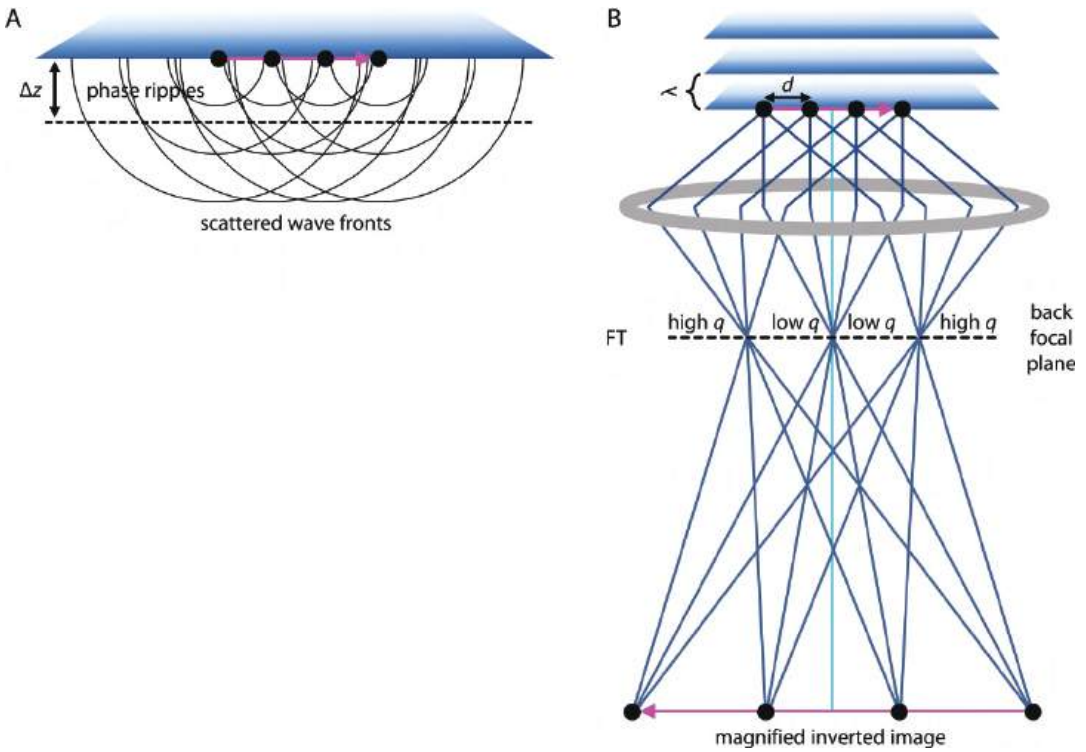


FIGURE 26.11: Scattering and interference in TEM. A: A continuous object (magenta arrow) is approximated by a series of scattering centers (black dots). Each scattering center disturbs the incoming plane wave (blue). The result is shown as concentric circles or phase ripples. All waves emitted from all scattering centers interfere with each other and with the incoming beam to different extents, depending on the defocus value Δz. Positive interference occurs at common intersection points of the wave fronts while at all other points, negative interference dominates. The chosen Δz determines which interference pattern of the waves is recorded on the detector. B: This figure is an enlargement of the top part of Figure 26.10. The plane wave shown in blue with "spacings" of wavelength λ interacts with a continuous object (magenta arrow). Four scattering centers spaced by a distance d, indicating the desired resolution to which the object should be imaged, are marked as black dots. The elastically scattered waves constructively interfere with each other in certain directions, giving rise to new wave fronts or "beams", while at the same time canceling each other in other directions. Unscattered electrons travel in the direction of the optical axis (cyan). The combination of all beams leads to a diffraction pattern in the back focal plane of the objective lens, which is the Fourier transform (FT) of the object.

26.4.2 Image contrast, amplitude contrast, and phase contrast

The goal of any microscopic technique is to obtain the best image possible, *i.e.* an undistorted high-resolution image with high contrast. *Image contrast* is defined as the intensity difference between the brightest and the darkest pixel in an image on the detector, normalized by the average intensity of all pixels of the image. The larger the image contrast, the higher the signal-to-noise ratio. Any image contains both amplitude and phase information, so image contrast is a combination of amplitude contrast and phase contrast.

In optical microscopy (Section 25.1), electromagnetic waves experience a sizeable change in amplitude upon interaction with objects: their amplitude contrast is large and may be detected directly. Hence, image generation in most optical microscopy techniques is dominated by amplitude contrast. Electrons, on the other hand, interact weakly with matter and TEM images have therefore very low image contrast in general. The image contrast in high-resolution single-particle TEM is particularly small because a relatively large objective aperture of ca. 100 µm has to be used to allow electrons with large scattering angles to pass (Figure 26.7). Large apertures also allow background scatter to pass, so inevitably reduce the image contrast. As a result, the molecular objects imaged in TEM are often barely visible in front of the background of a micrograph (Figure 26.6). Thin samples containing biological objects appear almost transparent to electrons. Transparency means that the

amplitudes of the scattered electron waves do not change much as they pass through the sample, which results in very small amplitude contrast. Therefore, image generation in EM necessarily relies on phase contrast, generated by interference of phase-shifted electron waves. In this respect, EM is a type of phase-contrast microscopy, a collective term for techniques that convert phase shifts in electromagnetic radiation passing through a transparent sample into recordable amplitude changes. It should be noted that in EM the phase shifts experienced by electron waves scattered by biological objects are not great either because biological objects are composed of light elements with small atomic numbers. Accordingly, such objects are called *weak phase objects*. Nevertheless, the small phase shifts and the ensuing phase contrast contain the majority of the signal available from EM.

To maximize phase contrast, which later must be converted into amplitude contrast for detection, we need to take a look at how strong the interference of scattered electron waves is, and how interference depends on the scattering angle θ and the focus of the incoming beam. When electromagnetic waves are scattered elastically, they always undergo a phase shift of $\pi/2$ or 90° relative to the incident beam. This phase shift happens with any radiation, and we have encountered it during scattering of X-rays and neutrons (Section 23.2.1). In addition to the phase shift of $\pi/2$ inherent in elastic scattering, the interaction of electrons with the Coulomb potential of the object introduces a smaller phase shift $\Delta\phi$. It is this additional phase shift that contains the information about the structure of the object. In other words, we have two sources of phase shifts, one caused by the elastic scattering process ($\pi/2$) and one caused by the structure of the object ($\Delta\phi$). The more strongly a wave is scattered, the larger $\Delta\phi$ becomes. $\Delta\phi$ increases with increasing scattering angle θ, and we recall that the scattering angle increases when the Coulomb interaction between incident electrons and the atomic nuclei is strong, which happens when electrons pass close to the nucleus (Figure 26.1).

For small scattering angles θ (at low resolution) where $\Delta\phi$ is essentially zero the electron waves experience an overall phase shift of approximately $\pi/2$ only. These low-resolution electron waves are almost parallel to the incident beam ($\theta \approx 0$) but completely out of phase. They will therefore show destructive interference with the incident beam, leading to elimination of most low-resolution information in the focal plane of the image. The negative interference of scattered waves at small scattering angles thus acts equivalently to a high-pass filter, which blocks low frequencies and allows only high frequencies (high-resolution information) to pass. An object imaged at low resolution under these conditions may indeed be invisible. An object imaged to high resolution under these conditions lacks the low-resolution information and is biased (distorted) toward high-resolution features. The reason for why the high-resolution features do not vanish in the same way as the low-resolution features do has two origins. One is the increasing contribution of $\Delta\phi$ with increasing θ to the total phase shift, which will deviate more and more from $\pi/2$. In addition, the scattered electron waves will no longer be (approximately) parallel to the incident beam. Both the deviation of the total phase shift from $\pi/2$ and the non-parallel nature of the scattered electron waves lead to increasing levels of constructive interference with the incident beam. In summary, the different extent of constructive interference with the incoming beam as a function of θ is the reason why phase information in the scattered beams is retained in the focal plane at higher resolution (larger q frequencies), but not at low resolution.

The problem of low-resolution information loss at focus may be solved in two ways, by using special optical components or by changing the focus. Optical components known as *phase plates* back-shift the phase of all scattered electron waves relative to the incident beam by $\pi/2$. The back shift reduces negative interference with the incident beam, preserving the signal at low resolution. Phase plates are useful to obtain a clear image on a detector for viewing and judging of the sample quality. However, they are challenging to align and may introduce additional aberrations into the imaging system. For high-resolution measurements, many modern electron microscopes do without phase plates and instead solve the problem by changing the focus. A change in focus of the incoming electron beam to a point away from the first intermediate image plane (Figure 26.4) is achieved by adjusting the defocus Δz of the objective lens, usually to underfocus. Defocus changes the phase relationships between scattered electrons and the incoming beam. The changed interference conditions avoid negative interference at small scattering angles θ and phase information (*phase contrast*)

is now converted into amplitude information (*amplitude contrast*) for the low-resolution range. A downside of recording images out of focus is blurring, which is equivalent to losing high-resolution information. A compromise must be struck between rescuing the low-resolution information and retaining the high-resolution information. We will see in Section 26.4.3 why EM data are usually collected at various defocus values to gain maximum phase contrast across the whole resolution range.

Now that we know how to maximize phase contrast in principle, we need a way to best convert it into measurable amplitude contrast.

26.4.3 Contrast transfer function

The first images collected in an EM experiment will be blurry and have small image contrast. As the main source of information in EM comes from phase contrast, enhancing the initially small image contrast requires optimization of the phase information in the images. The *contrast transfer function* (CTF) is a central concept in EM that describes how phase contrast information is optimized and then converted into amplitude contrast in the final images. The CTF depends on both instrument settings and the sample. Therefore, the parameters of the CTF need optimization during image processing to gain as much information from the scattered electrons as possible.

To understand the CTF and its optimization, we first return to the point-spread function (PSF). In optical microscopy, we introduced the PSF as the image of a point source obtained by a microscope (Section 25.1.3). The PSF thus contains all distortions of the object due to the imaging procedure and imperfections of the microscope. The image obtained can be described as the convolution (Section 31.16) of the object with the PSF:

$$image = object \otimes PSF \qquad \text{eq. 26.17}$$

We encountered the convolution before, in measured fluorescence decays, which are convolutions of the instrument response and the actual fluorescence decay (see Section 21.5.9, eq. 26.17), and in X-ray crystallography, where the crystal is described as a convolution of a lattice and the molecular transform (Section 24.4.2). The reciprocal image or phased diffraction pattern in the back focal plane of an electron microscope is the Fourier transform of the image. The Fourier transform of the convolution of two functions is the product of their Fourier transforms (Sections 31.15, 31.16). Following this rule, the Fourier transform of the convolution of the object with the PSF is the Fourier transform of the object multiplied by the Fourier transform of the PSF:

$$rec.\,image = FT\left(image\right) = FT\left(object\right) \cdot FT\left(PSF\right) = FT\left(object\right) \cdot CTF^{*} \qquad \text{eq. 26.18}$$

The Fourier transform of the PSF is related to the CTF. Specifically, the CTF is the complex conjugate of the Fourier transform of the PSF:

$$CTF = e^{-2\pi i \cdot FT\left(PSF\right)} \qquad \text{eq. 26.19}$$

or

$$FT\left(PSF\right) = Re\left(FT^{-1}\left(CTF\right)\right) \qquad \text{eq. 26.20}$$

The procedure to increase the image contrast now becomes apparent: take a crude EM image, Fourier-transform it into reciprocal (q) space, estimate the CTF and optimize its parameters, then apply the optimized CTF and back-transform into an image with improved contrast. This procedure is done for each micrograph.

We will see shortly how the initial CTF parameters are estimated from the Fourier transform of the EM image, which is also known as the *power spectrum* (Figure 26.12).

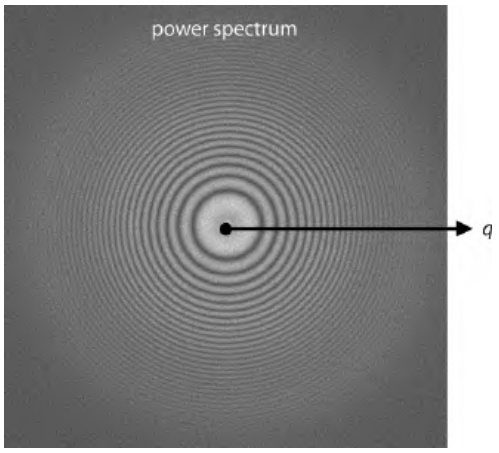


FIGURE 26.12: Power spectrum. Intensity distribution of the scattering pattern from amorphous carbon as a function of spatial frequency q. Amorphous carbon was used as the object to highlight the instrument features by minimizing (ordered) sample contributions. The result is the characteristic Thon rings, *i.e.* alternating intensity (amplitude contrast) with respect to the spatial frequency q. The contrast is large (white) at low resolution (small q) and changes periodically between small and large contrast at higher resolution (large q). The radius of the central region, the first white disc in the power spectrum, depends on the focus of the microscope. Instrument non-idealities, including mis-alignment or incomplete corrections of lens aberrations, would give a distorted power spectrum.

The power spectrum of an EM-generated image shows the amplitude contrast as a function of spatial frequencies q. Even if no actual object was present and the image is "empty", the distribution of amplitude contrast in a power spectrum is not uniform but shows a characteristic pattern of *contrast reversals* visible as bright and dark rings. The bright rings, known as *Thon rings*, are generated by positive interference of the scattered electron waves. The dark rings separating them are q-ranges where all scattered rays have canceled due to negative interference. The shape of the power spectrum gives valuable insight into the instrument settings. If the microscope is well-tuned, the Thon rings are evenly bright, circular, and sharp throughout. The signal in a power spectrum decreases with increasing q because fewer and fewer electrons are scattered at larger angles. The highest q at which a Thon ring can be discerned from the neighboring dark areas determines the maximum resolution attainable with the current instrument settings. The power spectrum thus aids in determining the highest useful recordable spatial frequency. In accordance with the Nyquist sampling theorem, the Nyquist frequency of the detector should be set to at least twice the highest spatial frequency q visible in the power spectrum (Section 26.1.4).

In the beginning of this Section we stated that the PSF contains all distortions of the object due to the imaging procedure and imperfections of the microscope, and that the PSF can be modeled by the help of the CTF. Thon rings visible in the power spectrum of an EM image are a manifestation of such image distortion, and they are specific to the EM image generation process. The oscillating amplitudes in the power spectrum have their origin in oscillating phase contrast. It is the task of the CTF to best describe the oscillating phase contrast introduced by the imaging system. The pattern of Thon rings is analyzed to estimate initial CTF-relevant parameters, including defocus, spherical aberration, and astigmatism. Although these parameters are known in the outset, they may drift during the course of the experiment. Importantly, the sample also influences the CTF by variations in electron density in the field of view, either across the object itself or due to differences in sample thickness. The parameters therefore need adjustment by mathematical fitting of the CTF with these parameters as variables. An optimized CTF is generated that better replicates the observed oscillation of contrast, *i.e.* the Thon rings, in the power spectrum. The new CTF is then used to correct the "reciprocal" experimental image in Fourier space. The correction seeks to compensate for the oscillation in phase contrast and to enhance amplitude contrast. The CTF-corrected "reciprocal" image is then inverse Fourier transformed into a corrected real-space image with increased contrast and resolution.

We will now turn away from CTF optimization and take a look at the mathematical form of the CTF to see how changes in its parameters influence EM image contrast. We saw in Section 26.4.2 that an EM image taken in focus is nearly invisible due to efficient negative interference of scattered electron waves and incident beam. Defocusing the imaging system changes the interference conditions and produces a visible image. One way to introduce defocus is to change Δz, but what is a good value for Δz? A second parameter that introduces defocus is the spherical aberration C_s, usually a constant for any electromagnetic lens and specific for a certain microscope. C_s also helps increasing the phase contrast. The two parameters, C_s and Δz, both determine the phase shift and are central to the CTF. Typical values of C_s for uncorrected electron microscopes operating at of U_{acc} = 200 kV are between 0.5 mm and 2 mm. The phase shifts at wavelength λ due to the spherical aberration C_s and the chosen defocus value Δz combine into a single phase factor χ given by

$$\chi(\boldsymbol{q}) = phase\ shift = \frac{\pi \cdot C_s \cdot \lambda^3 \cdot q^4}{2} + \pi \cdot \lambda \cdot \Delta z \cdot q^2 \qquad \text{eq. 26.21}$$

$\boldsymbol{q}$ is the reciprocal space vector leading to the observed spatial frequency q (Section 26.4.1). It is a vector because EM images are two-dimensional and points in the images are best described by a vector of coordinates (q_x,q_y). The magnitude q of the scattering vector $\boldsymbol{q}$ in electron microscopy is sometimes referred to as f or u, with the vectors $\boldsymbol{f}$ or $\boldsymbol{u}$, but for sake of continuity, we will continue to use q and $\boldsymbol{q}$. The strong dependence of χ on q in eq. 26.21 reflects that the phase shift of scattered waves increases with the scattering angle. The phase shift is not only a function of θ (or $\boldsymbol{q}$), but also of the density distribution of the object after projection into two dimensions. Therefore, different parts of a specimen will generate different phase shifts and hence maximum phase contrast under different defocus values Δz. Likewise, different defocus values enhance phase contrast at different resolutions. The result is that each defocus value gives rise to a specific phase contrast transfer function that describes the phase contrast of an object as a function of q. The ideal CTF for coherent scattering is defined as the sine of the phase factor χ:

$$CTF_{ideal}(\boldsymbol{q}) = \sin(\chi) = \sin\left(\frac{\pi \cdot C_s \cdot \lambda^3 \cdot q^4}{2} + \pi \cdot \lambda \cdot \Delta z \cdot q^2\right) \qquad \text{eq. 26.22}$$

This function is negative for positive phase contrast, which images atoms dark against a bright background, and positive for negative contrast (bright against dark background). Being a sine function, the CTF oscillates between values of –1 and +1, periodically inverting the signal (contrast reversal) to describe the Thon rings. At its zero points, $CTF_{ideal}(\boldsymbol{q}) = 0$, there is no contrast and hence no image information at all for this spatial frequency. The sinusoidal nature of the CTF (and the Thon rings) explains why it is physically impossible to image an object with maximum phase contrast over a large resolution range. Especially around the zero points of the CTF there will be information loss due to small phase contrast. Nothing can be done about this effect on the phases, it can only be adequately described by CTF optimization as discussed above. However, what can be done is to tailor the experimental settings such that maximum phase contrast is achieved. If the object is imaged under different defocus values Δz, the zero points of the different CTFs (Figure 26.13) shift, and when the measurements are combined phase contrast will be available over the entire resolution range.

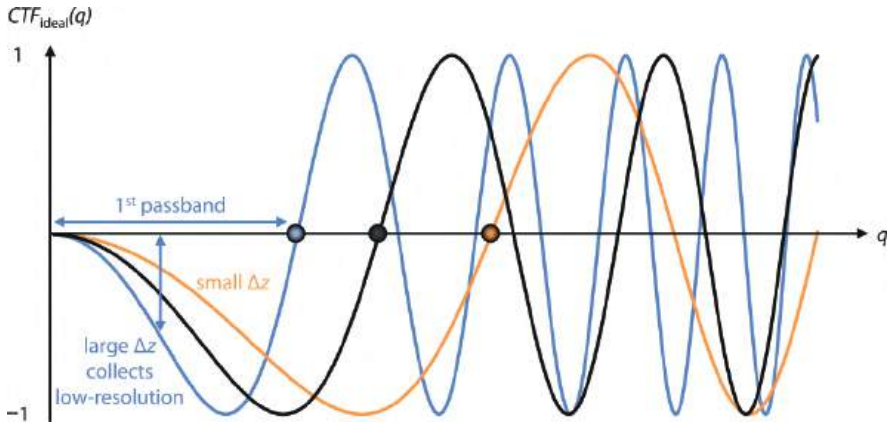


FIGURE 26.13: Contrast and underfocus. Ideal CTF curves for large (blue), intermediate (black), and small (orange) negative defocus values Δz. The more negative Δz, the more low-resolution signal can be collected, as seen by the blue curve with large Δz that becomes negative more quickly (vertical double arrow) than the other curves with smaller (less negative) Δz. The first zero point shifts to smaller spatial frequencies q with increasing negative Δz, so CTF with large Δz have more zero points within a q-range than CTF with small Δz. At any zero point there is no image information, hence data collection must be performed at various different defocus values. Typical defocus values used in cryo-EM data collection are between –500 nm and –1000 nm. The regions between each pair of zero points are called passbands. The aim is to make the first passband (indicated by a double arrow) as large as possible at maximally large defocus.

Large underfocus (more negative Δz) leads to increased phase contrast at a certain small q (Section 26.4.2), but also to a higher oscillation frequency of the CTF, with more information-free zero points. On the other hand, smaller underfocus (less negative Δz) gives fewer zero points, but also less contrast at low resolution. For an optimal signal transfer from object to magnified image the number of zero points should be small and at the same time the contrast should be large and negative over a broad band of q, *i.e.* the first *passband* or q-region of signal transmission should be as large as possible. This condition is met at the *Scherzer defocus* or *Scherzer underfocus*:

$$\Delta z_{\text{Scherzer}} = -2 \cdot \sqrt{\frac{C_s \cdot \lambda}{3}} \approx -1.15 \cdot \sqrt{C_s \cdot \lambda} \qquad \text{eq. 26.23}$$

The Scherzer defocus may be derived from eq. 26.22 and depends only on two microscope properties, the spherical aberration C_s and the wavelength λ (determined by the acceleration voltage U_{acc}). At $\Delta z_{\text{Scherzer}}$, the slope of the CTF is small over extended regions of q (Figure 26.14), *i.e.* the first passband is broad. By balancing the effects of the spherical aberration and defocus terms in eq. 26.22, $\Delta z_{\text{Scherzer}}$ maximizes at the same time the phase contrast of a weak phase object and the resolution of the data.

Figure 26.13 shows ideal CTF curves that have the same amplitude over all q. In practice, the higher spatial frequencies q of the CTF become ever smaller in amplitude, limiting the high-resolution data in a similar fashion as the drop-off in intensity of diffracted X-ray beams. A damping effect intrinsic to an electron microscope is responsible for the loss in signal intensity at higher resolution. Multiplying the damping factor with the ideal phase $CTF_{ideal}(\boldsymbol{q})$ gives the *effective* $CTF_{eff}(\boldsymbol{q})$ (Figure 26.14). This damping is caused by shortcomings of the electron microscope and the sample, such as angular spread of the source, chromatic and spherical aberrations, and detector imperfections, as well as specimen drift and vibration. The overall effect of the damping factor is a suppression of phase contrast at large spatial frequencies, *i.e.* high-resolution information is lost.

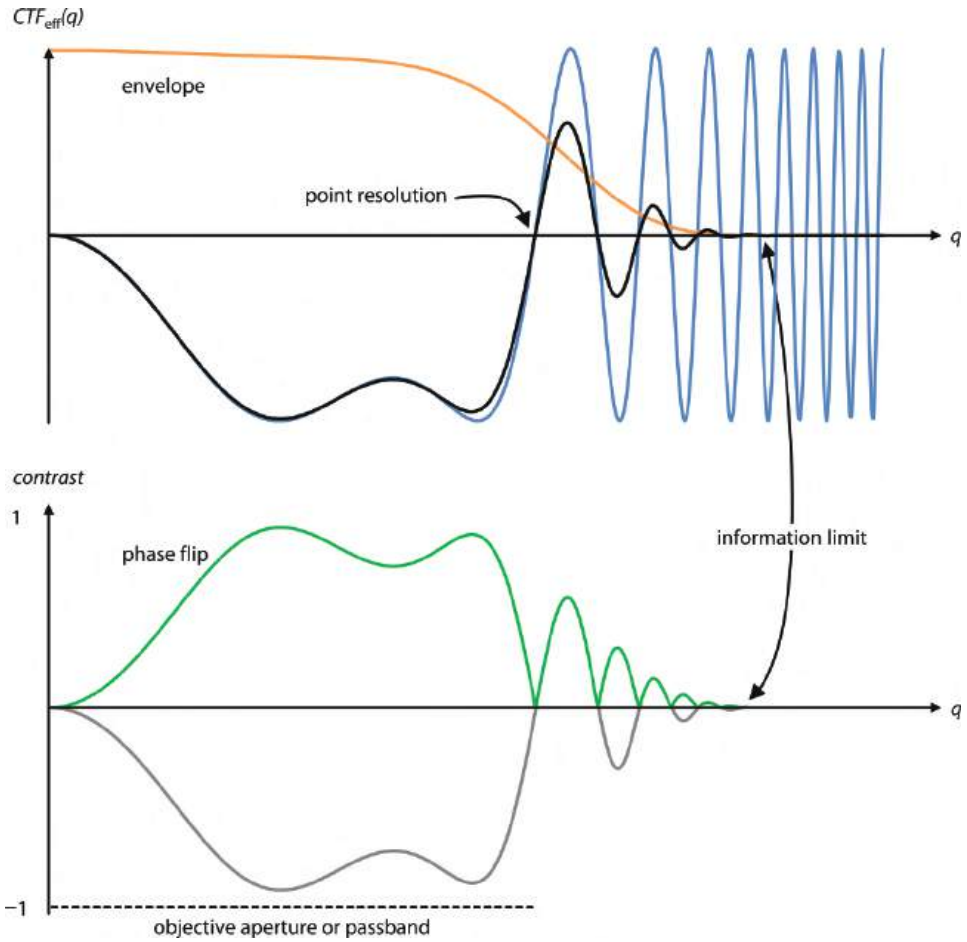


FIGURE 26.14: Effective CTF. Multiplying the ideal CTF (blue) with a function describing the damping effects (orange) results in a damped signal at higher spatial frequencies (black). The CTF calculated at the Scherzer defocus has a maximally broad passband that increases the range of spatial frequencies with high transmittance, allowing more phase signal to reach the image plane.

When a microscope operates at the Scherzer defocus (eq. 26.23) and images a very thin sample, the first zero point of the CTF has a special name, the *point resolution* d_0, given by the *Scherzer expression* (eq. 26.24).

$$d_0 = 0.66 \cdot C_s^{1/4} \cdot \lambda^{3/4} \qquad \text{eq. 26.24}$$

The point resolution is the smallest detail that can be interpreted directly from a micrograph. This value increases with increasing acceleration voltage (which determines the wavelength λ) and decreasing spherical aberration C_s. Any higher resolution details need additional image processing. The highest spatial frequency at which the damping function, and hence CTF_{eff}, become zero is termed the *information limit* (Figure 26.14). At this somewhat arbitrarily chosen frequency, no useful phase information is left, and it is this frequency that is used as an indication for the highest resolution of a TEM image after its CTF_{eff} function has been found by the CTF-optimization algorithms.

26.5 THREE-DIMENSIONAL ELECTRON MICROSCOPY

The micrographs recorded by an EM are two-dimensional projections of three-dimensional objects. The object can be reconstructed in three dimensions when imaged from many different angles of view. In electron tomography and crystallography, different angles of view are achieved by changing the tilt angle of the sample. In single particle cryo-EM, this is achieved by depositing particles in random orientation on a grid, such that different two-dimensional projections of the object are obtained by imaging the entire grid. The result of any three-dimensional reconstruction from electron micrographs is a three-dimensional representation of the electrostatic potential of the nuclei of the atoms in the object, the *potential map*. The potential map has features similar to the electron density map generated by X-ray diffraction, but the physical origin of the signal is different: while in X-ray diffraction the elastic scattering of photons by electrons reports on the density of electrons, in EM it is the potential (positive charge) of atomic nuclei that is probed by the elastic scattering of electrons.

In the following Sections, we will discuss imaging of three-dimensional objects from series of projections obtained by cryo-electron tomography (Section 26.5.1) or single-particle cryo-EM (Section 26.5.2). EM imaging by tilting of ordered samples, such as two-dimensional crystals of membrane proteins and three-dimensional nanometer-sized crystals, is described as part of crystallography in Section 24.8.

26.5.1 Cryo-electron tomography

In cryo-electron tomography, $\leq$ 500 nm-thin sections of a biological specimen, kept at liquid nitrogen temperature (77 K), are imaged at a range of tilt angles up to $\pm$70° in steps of 0.5–5°, either around a single axis or around several axes. Different projections in such a *tilt series* are collected under considerations of maximal phase contrast but minimal radiation damage, then carefully aligned and combined into a three-dimensional image of the object.

A complete dataset consists of tilt series covering tilt angles in the range from –90° to +90°. Unfortunately, not all of these tilt angles can be realized due to technical limitations of the microscope and by the sample. The stage holding the sample in the microscope may typically be tilted by only $\pm$70° before it obstructs the incoming electrons (from the z-direction). As a result, a wedge of volume data in EM tomography is always missing along the z-axis that is perpendicular to the sample in the xy-plane. If the sample can be tilted about more than one axis, combining the data reduces the missing wedge to a cone. Still, there is a systematic part of data missing along the z-direction with two consequences. First, the reconstruction is of lower resolution in the z-direction. Second, the three-dimensional reconstruction will be elongated in the z-direction, called the *missing wedge artifact*. The elongation is an effect of the anisotropic sampling of data and the Fourier transformation. The sample itself also limits the information content achievable from high tilt angles. At tilt angles near the maximum allowed by the microscope, the path lengths of the electrons through the sample are much increased compared to low tilt angles. The larger the path, the more electrons are absorbed, inelastically scattered, or experience multiple events of elastic scattering. All of these events reduce the signal at higher tilt angles. As a consequence, volume data from higher, but achievable, tilt angles is incomplete, further limiting the resolution of the reconstruction in the z-direction (Figure 26.15).

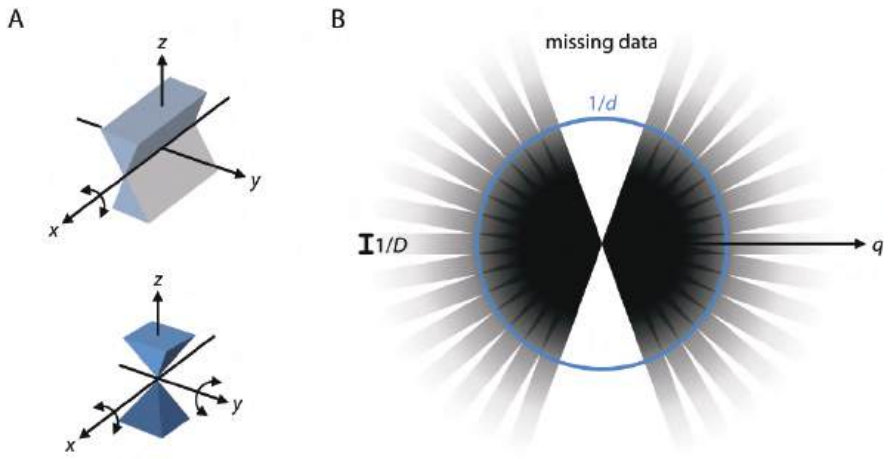


FIGURE 26.15: Tilt series and projections in reciprocal space. A: Missing data from instrument-limited tilt angles. The top shows a missing wedge in a series tilted about a single axis, the bottom a cone from a series tilted about two perpendicular axes. B: Simplified visualization of Fourier-transformed projections from different tilt angles in reciprocal space. Projections of thickness *D* in real space leads to slice planes of "thickness" 1/*D* in reciprocal space that overlap at small frequencies (low resolutions, small frequencies *q*, toward the center of the figure). The blue circle denotes the frequency $q = 1/d$ where the slices overlap to fully describe the object over 360°. Regions where the signal is augmented too much by the overlap (dark spikes) need downweighting by filtering for a realistic reconstruction. The signal drops for larger frequencies, indicated by the fading grey tone. Larger frequencies suffer from information-free areas (white space), small ones between neighboring projections and larger ones at very large tilt angles. The small missing data areas can only be completed by inclusion of projections corresponding to tilt angles filling these gaps. The large missing wedges at the top and bottom cannot be filled from a single tomogram.

Missing wedges in tomogram datasets are a severe problem. In crystallography, missing data wedges are also common but can often be avoided with high-symmetry crystals that fill the missing wedges with data by symmetry-related reflections (Section 24.4.3). This is only an option in EM if the object studied is of high symmetry, for example microtubules, flagella, or filamentous viruses. Sometimes, the missing data parts in tomography may be filled by including data from an object with a different orientation in the *xy*-plane or by *sub-tomogram averaging*. If the object embedded in the sample is present multiple times and in various orientations, the potential density in these sub-volumes of the tomogram can be placed into a common coordinate system, *i.e.* superposed, and then averaged to improve the signal-to-noise ratio. The results of sub-tomogram averaging can be spectacular: Tomography normally reaches resolutions of about 0.9 nm. With sub-tomogram averaging completing the data and increasing the signal-to-noise ratio of the data, structures of 0.27 nm resolution have been determined.

Radiation damage is a perpetuous problem in EM (Section 26.2.3) and especially severe in tomography with the comparatively thick samples that require large electron doses for transmission imaging. Large path lengths at high tilt angles favor radiation damage. Therefore, in tomography images are first collected from small tilt angles. In the absence of damage it would also be expected that the phase contrast at angles $\pm\theta$ is about the same, but if images at these angles are not generated close in time, their phase contrast will be different due to different extents of radiation damage at $+\theta$ compared to $-\theta$. As radiation damage cannot be avoided, data for $\pm\theta$ is collected close in time using a dose-symmetric tilt-scheme: First collect data at small tilt angles where the path length through the sample is comparatively small, then proceed to larger tilt angles (tilt-scheme *e.g.* ±3°, ±6°, *etc.*). The total dose should not exceed 8000 e nm^{-2} (80 e $Å^{-2}$), which is facilitated by setting the microscope to lower magnifications to spread the dose over a larger field of view. Nevertheless, as soon as radiation damage sets in, the image data become anisotropic and so does the resulting reconstruction.

Individual two-dimensional projections from a tilt series then need to be aligned for construction of the three-dimensional potential map/image. Precise alignment of the tilt images corrects for tilt-induced distortions and shifts and is a prerequisite for extracting high-resolution information. The process is non-trivial due to the low phase contrast in the projections. Fiducial markers can help in alignment. These are either nanometer-sized electron-dense gold particles embedded in a specimen or "internal" marks such as ribosomes or other comparatively electron-dense particles present in the sample that produce locally high phase contrast. Tracking the positions of the fiducial markers in a tilt series allows accurate calculation of the rotation and translation operators between the images, with which a precise superpositioning of the images is calculated. If no fiducial markers are present, tilt series must be aligned by *patch tracking*: instead of following a spot-sized, high-contrast fiducial marker, a more spread-out pattern of low contrast serves as the marker. In this method, an image is sub-divided into overlapping sections, and the movement of the sections is monitored over the tilt series by cross-correlation. Patch tracking is less accurate than tracking of fiducial markers, especially because the phase contrast of images keeps decreasing with increasing tilt angles, and thus the correlation between patches in images taken at different tilt angles diminishes.

Once the two-dimensional projections are aligned, a three-dimensional potential map is calculated. The three-dimensional reconstruction is achieved by back-projection of the aligned projections into a common volume. Back-projection can be conducted either in real space or in reciprocal (Fourier) space. While a reconstruction in real space is intuitive and allows straightforward application of filters, including smoothing (low-pass filtering), sharpening, edge detection, and noise reduction, it is computationally inefficient. Most three-dimensional reconstructions are done in Fourier space, because applying these filters is a convolution operation, and convolutions in Fourier space are a matter of simple multiplication. A second advantage of reconstructing in Fourier space is the exploitation of the *projection slice theorem* (Box 26.1). Instead of directly performing a three-dimensional Fourier reconstruction of the entire object, the projection slice theorem greatly simplifies the computation by separating the three-dimensional reconstruction into a series of two-dimensional slice reconstructions.

BOX 26.1: PROJECTION SLICE THEOREM AND ITS USE IN EM.

For better handling of experimental noise and CTF-correction, the back-projection is often done in Fourier space or reciprocal space using the properties of the projection slice theorem, also known as central slice theorem or Fourier slice theorem. This theorem states that if any n-dimensional function is projected into n–1 dimensions and then Fourier-transformed, the result is identical to taking first the n-dimensional Fourier transform of the function and then slicing the result through its origin at the angle of the projection. For EM this means that the Fourier transform of the two-dimensional projection of a three-dimensional object represents a central slice through the three-dimensional Fourier transform of the object with the angles of the slice plane given by the rotation angles of the projection (Figure 26.16).

The Fourier transforms of the two-dimensional projections are combined to reconstruct slices through the three-dimensional Fourier transform of the object. The reconstructed slices are then inverse Fourier transformed to obtain slices in real space, which are then combined into a volumetric three-dimensional reconstruction.

(*Continued*)

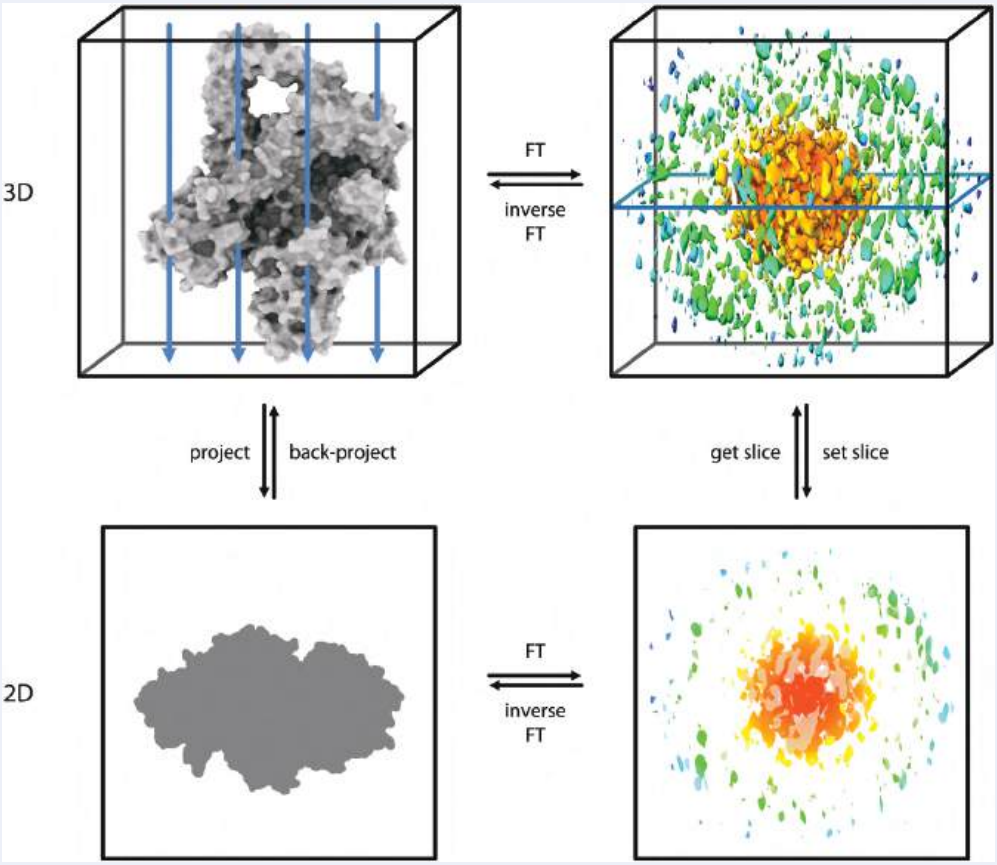


FIGURE 26.16: Real and reciprocal space relationships in cryo-EM. If a three-dimensional object is projected into a plane and then Fourier-transformed (left-hand side of the Figure), the result is identical to first Fourier-transforming the object and then extracting a slice. This relationship proves useful in calculating the back-projections from both tomograms and single particles (Section 26.5.2).

26.5.2 Single-particle cryo-EM

An alternative to tilting of the sample for collection of images from all possible angles is to distribute single objects on a grid in as many random orientations as possible. The grid is then imaged from a single view, such that all projections of the object are collected in a single micrograph. Identification of projections, their sorting into classes of different orientations, and back-projection into space then allows the three-dimensional reconstruction of the object. This mode of data acquisition is the basis of high-resolution single-particle EM. In recent years, single-particle EM has produced three-dimensional reconstructions of large and often asymmetric biological macromolecules with resolutions of < 0.2 nm (2 Å). Examples include a 1.08 MDa sub-complex of the human spliceosome (PDB-ID 7q4o, 2.1 Å), the 2.6 MDa yeast fatty acid synthase (PDB-ID 8prw, 1.9 Å), and the human 80S ribosome (PDB-ID 8glp, 1.67 Å).

The first step in the analysis of single-particle cryo-EM images is the identification of particles in the projections (*particle picking*). The term *particles* describes any projections of the object that may be distinguished from background noise. The micrographs are analyzed for particles by cross-correlation with a set of calculated projections from a reference. The reference is a previously determined structure (compare molecular replacement in Section 24.4.4.3), a set of manually (by eye) selected projections, or a simple shape with low-resolution features of 3–4 nm. The micrograph is centered on each pixel and a two-dimensional box around the pixel is cross-correlated with all

calculated projections of the reference. A possible projection of the molecule (particle) is identified by the resulting large correlation coefficients over a contiguous area of pixels.

Just as in molecular replacement in X-ray crystallography, model bias is a problem in single particle EM: if the reference projection is calculated with a high resolution, the correlation with the experimental data will be high as well, leading to false-positive identification of particles on the micrograph. Removing the high-resolution details from the reference by low-pass filtering helps reduce model bias in particle finding. Machine learning algorithms (Section 20.2) that have been trained to classify noise, de-noise images, or recognize anomalies in micrographs are increasingly being used to extract small particles, *i.e.* the projections of small (difficult) objects from the noisy micrographs. However, objects of molecular mass < 100 kDa will usually not produce sufficient contrast and hence cannot be identified from the micrographs. The limitation can be overcome by artificially increasing an object's molecular mass with a specifically binding antibody Fab (50 kDa) or another large binding partner and determining the structure of the larger complex.

Sorting the particles identified from the micrographs into classes of different two-dimensional projections of the object is the next step. For this *2D classification* the sub-images of the particles are extracted from the micrograph, rotated in the plane of the micrograph for maximum correlation, and sorted into classes. Each class corresponds to a certain projection, *i.e.* the same original orientation of the object. 2D classification can be straightforward for highly symmetric objects such as viruses, fibrils, and homo-oligomers (Figure 18.38), but it is usually more difficult for asymmetric objects for which many more orientations exist than for symmetric objects. Distinguishing between projections of a single conformation from slightly different angles and various conformations from the same projection angle is also difficult. A compromise is needed to keep conformations separate while retaining a high enough number of particles in each class. Modern alignment algorithms rely on maximum likelihood and Bayesian statistics to automatically sort the projections into classes representing a single conformation at a certain orientation projected into two dimensions. Similar to the projection-identification step, the sorting process may be monitored by calculating the cross-correlation of the extracted projections with calculated projections of a reference. However, if a reference is unavailable, a bootstrapping method known as *principal component analysis* (PCA) may be used to separate the projections into major 2D classes of molecular orientations. PCA is a statistical method often used in image analysis that finds subsets of "correlated data" points in raw data. In TEM, the correlated data are the projections belonging to similar molecular orientations, and the raw data are the noisy micrographs.

Following the first round of 2D classification, the projection sub-images in each 2D class are averaged. The process of identifying and sorting of projections from the micrographs is then repeated iteratively, using the class average image as a new reference. The iteration is stopped when scoring functions such as cross-correlation or maximum likelihood (see Section 24.4.6) do not increase any more, or the maximum number of classes chosen by the user has been reached. During the process of iterative 2D classification, low-populated classes of just slightly different orientations may be generated. Such classes represent slightly different conformational states of the object, and especially when the object shows high conformational variability there can be many of them. Rejecting low-populated classes bears the risk of ignoring potentially useful data but for high-resolution structure determination, it is important to have fewer, conformationally homogeneous and highly populated classes. To reach this goal with more flexible objects, binding of an antibody Fab can reduce the conformational entropy of the complex, which in turn can reduce the number of 2D classes with the promise of higher-resolution reconstructions.

After sorting the projections into 2D classes, the next step is to correlate which three-dimensional orientation of the imaged object led to this two-dimensional projection. This operation puts all projections into the same reference system by way of rotation matrices or Euler angles for the following back-projection into a three-dimensional object. In the beginning, these rotation angles are unknown, hence they are first approximated and then iteratively refined. The starting angles are approximated with the help of a three-dimensional reference object. The object can be an *ab initio* model generated by modern algorithms such as CryoSparc, a low-resolution structure from a related

protein, or a low-resolution three-dimensional image of the object calculated by averaging of all two-dimensional projections found on the micrographs using random orientation angles. In the first step of the iteration, new two-dimensional projections are calculated from the reference object under all possible angles and these calculated projections are compared with the measured projections from the micrographs. Once a fit of a calculated and measured projection is found, the orientation angles of the calculated projection are assigned as the orientation angles of the measured projection. After this assignment is completed for all 2D classes, a first back-projection and three-dimensional reconstruction is calculated (Figure 26.17).



FIGURE 26.17: Back-projection for three-dimensional image reconstruction. The orientation angles of randomly oriented particles are determined and the projections (black) placed onto the surface of a sphere (blue). Back-projection of these images into a volume at the center of the sphere generates an average three-dimensional density map of the object. An even distribution of the projections over the sphere surface results in an undistorted reconstruction at the center. The figure shows the back-projection in real space, but in practice the calculation is performed in reciprocal space.

Quite analogous to the reconstructions in tomography, the calculations for single-particle reconstructions are more conveniently performed in Fourier space than in real space, using the advantages of splitting the calculations into smaller parts and the projection slice theorem (Figure 26.16). The first three-dimensional reconstruction typically is rather noisy and hence lacks high-resolution detail. It serves as a new reference model for another round of improvement of the orientation angles of each 2D class. New projections are calculated and the process of angle refinement and back-projection is repeated until the reconstruction does not improve further.

At this point, the information content of the reconstruction may still be unsatisfactory. Reasons include sample heterogeneity such as different conformations of single molecules, structural changes induced by cryo-cooling and radiation damage, and variations in stoichiometry of multimeric complexes that have not been accounted for during particle picking. PCA and maximum likelihood methods are again able to help by sub-classifying the first three-dimensional reconstruction into several three-dimensional shapes, or 3D classes. Ideally, these 3D classes represent individual conformations or complexes with different stoichiometry. Another round of 2D classification using the 3D classes as reference models can be done.

Depending on the quality of the data, the final result of a single-particle reconstruction may be a single high-resolution potential map or a set of lower resolution maps representing different conformations of the object. Which way the cryo-EM data are evaluated depends on the biological question to be answered. If the question is about molecular details of a single conformation, a high-resolution map is required. High-resolution structure determination requires highly populated 2D classes from all possible orientations. Each class must represent only a single conformation or high-resolution information is lost

in the reconstruction. This may require removing a substantial amount of data that does not fit the 2D class average: including the projections of different conformations in the same 2D class drastically decreases the signal-to-noise ratio. Including such blurry 2D classes in the reconstruction will blur the three-dimensional potential maps as well, *i.e.* high-resolution information is lost.

If, on the other hand, the question is to decipher the different conformational states of an object for insight into dynamics, folding, and possibly catalysis, lower resolution has to be accepted (and may be sufficient). Such three-dimensional reconstructions are generated either by including low-populated 2D classes or by splitting highly populated, but blurry, 2D classes into smaller ones that represent clear projections of slightly different conformations.

Optimal grids may provide enough particle views to achieve both goals from a single EM experiment, *i.e.* high-resolution reconstructions of several conformations. Usually, however, the opposite is true and the distribution of particles on the grid is not random. Non-random particle distribution leads to a systematic lack of data over an angle range, which introduces artifacts in the three-dimensional reconstructions similar to the missing wedge artifact in tomography (Figure 26.15). Preferential orientation of objects on the grid is especially prominent for elongated objects or membrane proteins embedded in nanodiscs (Figure 18.50). The best solution to this problem is preparing grids with no preferred orientation of the objects by adjusting the sample support, buffer conditions, and cryo-protection protocol. Alternatively, if the object under study has internal symmetry (*e.g.* the proteasome in Figure 26.17), a lack of random orientation angles can sometimes be compensated by applying this symmetry during data analysis.

26.5.3 Resolution and interpretation of cryo-EM potential maps

We used the term *resolution* of potential maps throughout this Chapter without exactly specifying how it is estimated. In crystallography, the practical resolution limit is nowadays often defined by calculating the half-dataset correlation coefficient $CC_{1/2}$ and setting the high-resolution limit when $CC_{1/2} \leq 0.3$ (Section 24.4.3). A similar approach, called the *Fourier Shell Correlation* (FSC), is used in cryo-EM. The (image) data is split in two arrays, and two independent three-dimensional potential maps are calculated by three-dimensional reconstruction. Calculation of the correlation between the two volumes in real space is difficult but becomes simpler in reciprocal space where the correlation can be written as a product. Hence the FSC calculation starts by Fourier-transforming the potential density maps into the frequency domain. The data points in the frequency domain are represented by spherical coordinates. The distance vector of a data point from the origin is the radial frequency q_r. FSC is calculated as a function of these radial frequency shells as

$$FSC(q_r) = \frac{\sum_{q_{i,r} \in q_r} F_1(q_{i,r}) \cdot F_2(q_{i,r})^*}{\sqrt{\sum_{q_{i,r} \in q_r} |F_1(q_{i,r})|^2 \cdot \sum_{q_{i,r} \in q_r} |F_2(q_{i,r})|^2}} \qquad \text{eq. 26.25}$$

F_1 is the structure factor for volume 1, F_2^* is the complex conjugate of the structure factor for volume 2, and $q_{i,r}$ is the individual voxel at radial frequency q_r. The sum with $q_{i,r} \in q_r$ means summation over all Fourier space voxels $q_{i,r}$ in the radial frequency shell q_r. A plot of FSC(q_r) *versus* the frequency shells q_r shows a sigmoidal shape (Figure 26.18). At a certain value of q_r where the FSC falls below a certain value, the overall resolution cut-off is set. Setting FSC = 0.143 as the lowest value at which data still correlate well enough to be used in three-dimensional reconstructions allows to compare resolutions of cryo-EM potential maps with X-ray crystallographic electron density maps. FSC only measures correlation of volumes, not resolution *per se*. The radial frequency at which FSC < 0.143 thus serves as a proxy for the practical high-resolution limit. The resolution limit calculated from the FSC *versus* q_r plot should be reflected by the features of the potential maps, *e.g.* at 0.3 nm resolution side chains should be visible in the maps (see Figure 24.27 for a crystallographic example). Conformational heterogeneity of the object damps the high-resolution features both in crystallographic electron density and in cryo-EM potential maps. Even though the criterion FSC < 0.143

defines a global resolution, the *local resolution* at different positions in a potential map can vary strongly. Local resolution is higher close to the well-ordered hydrophobic core of a molecule and decreases toward its surface (Figure 26.18).

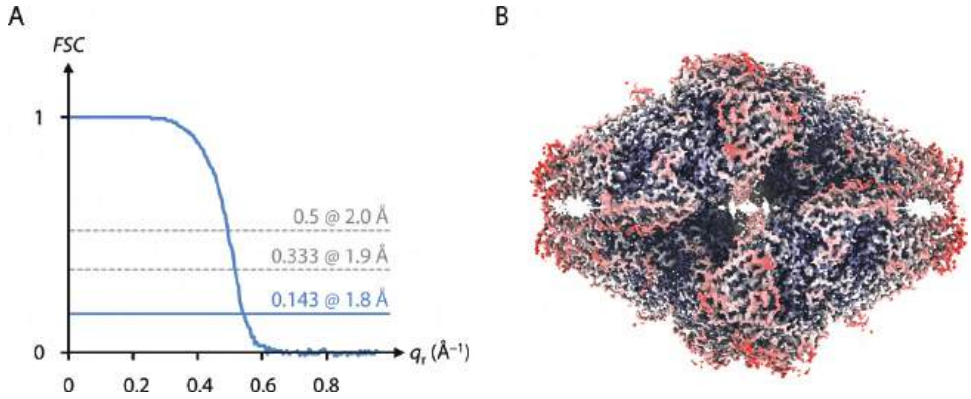


FIGURE 26.18: Fourier Shell Correlation and local resolution in EM maps. A: *FSC* as a function of spatial frequency q_r for β-galactosidase (EMDB entry 21995). *FSC* = 0.143 suggests a high-resolution cut-off of 1.8 Å. B: Surface representation of the potential map of the tetrameric enzyme colored blue, white, and red according to the local resolution ≤ 1.8 Å, between 1.8–2.1 Å, and > 2.1 Å. The resolution is highest in the center of the molecule and in depressions on the surface (blue color). At the perimeter of the molecule and especially at convex surface parts, the resolution decreases quickly (red color).

Potential maps are interpreted by atomic models. The models are built by manual or automated fitting of domains, secondary structure elements, main chains, residues, and rotamers, quite in the same way as crystallographic maps (Section 24.4.5). Potential density maps calculated from TEM data tend to show more local details than density maps from X-ray crystallography at the same nominal resolution. Potential maps of very high resolution (< 0.2 nm) enable automated *de novo* building of a model from the protein and/or nucleic acid sequence. For lower-resolution maps (0.3–0.4 nm), standard secondary structure elements such as helices, small β-sheets, and turns may be fit automatically, but model completion remains a manual task. Interpretation of maps with even lower resolution requires prior structural information, *i.e.* models from NMR or X-ray crystallographic studies. These structures are fit into the map as rigid bodies by a six-dimensional rotation and translation search (similar to placing of the model in molecular replacement; Section 24.4.4.3). In this way, a large, hetero-multimeric complex can be constructed by fitting the previously determined structures of individual constituents into the potential map. In many cases, a rigid-body fit alone does not fully explain all features in a cryo-EM map. In such cases, conformational changes in the rigid bodies need to be inferred, and the model needs manual adjustment. It is often these conformational differences from rigid bodies that contain the most useful information in cryo-EM structures.

The initial structural model then needs further refinement. At resolutions < 0.4 nm, real-space refinement of the initial model into the map adjusts the coordinates of the model to fit the map with a maximum real-space correlation coefficient, which is calculated on a per-residue basis. Models for lower-resolution maps (> 0.4 nm) may be improved using molecular dynamics (Section 20.1), which allows simplifying the atomic models by coarse-graining (Section 20.1.4) or hinge movements of sub-domains. Both approaches reduce the number of parameters to be refined. Also, additional restraints from independent experimental data can be included in a hybrid approach (see Box 20.1) to improve the model. The final step is validation of the structural model by much the same methods as are used for crystallographic models (Section 24.4.6), including the Ramachandran plot and real-space density correlation coefficients. The best-interpretable map generated from the data and the atomic model interpreting it are deposited in the *EM databank* (Box 26.2).

BOX 26.2: EM DATABANK.

Electron density maps from EM experiments are deposited in the EM databank, (www.ebi.ac.uk/emdb/statistics) which held 39400 entries in mid-2024. Almost half of all entries are structural models of icosahedral viruses and ribosomes. The other half is somewhat dominated by molecules that display high internal symmetry, such as GroEL (see Figure 18.39), the proteasome, and apoferritin, which serves as a reference model and test case for EM setups. Single-particle cryo-EM has contributed the majority (83%) of all structures. Helical reconstructions, tomography, and subtomogram averaging contribute about 5–7% each. Electron crystallography currently contributes < 1% of all entries.

About 75% of all entries have resolutions < 0.4 nm, and 27% have a resolution < 0.3 nm, enabling detailed studies such as molecular interactions and catalytic mechanisms. The remaining 25% of all EM structures, with resolutions > 0.4 nm, do not contain detailed structural information. Nevertheless, they answer many biological questions about the composition and conformational plasticity of large complexes. With the recent developments in detector and image processing and the associated increase of high-resolution structures, more and more maps can be interpreted in terms of atomic models. These high-resolution models are also deposited in the PDB and cross-referenced, so the two databases interface with each other.

EM databank statistics show trends in the field of electron microscopy: structure determination of ever smaller objects of < 100 kDa molecular mass becomes increasingly possible. The structures of smaller molecules are often determined in complex with heavier binding partners. For example, binding of specific antibody Fab (Figure 24.5), which contributes 50 kDa, pushes the molecular mass of the complex into a range accessible to EM. As a result of the detector and processing software developments, 1–2 orders of magnitude fewer particles are needed for a high-resolution image reconstruction. Further advances in hardware and software will keep pushing the high-resolution limits of single-particle and tomography reconstructions. Improved sample preparation methods will enable *in situ* cryo-EM, which images biological specimens in their native cellular environment under near-physiological conditions. Automation of data collection and processing promises faster structure determination and higher throughput of cryo-EM for structure-based drug design. Additionally, deciphering dynamic processes and conformational changes in biological macromolecules should become more routine with improved classification algorithms based on machine learning.

QUESTIONS

26.1 Derive a formula that gives the relative error in % of the wavelengths of electrons, calculated with and without taking relativistic effects into account, as a function of the acceleration voltage. At what acceleration voltage does the relative error exceed 1%?

26.2 A thick specimen of about 100 nm is imaged in a TEM operated at 100 kV with opening angle of 4 mrad. The chromatic aberration is 3 mm and leads to an energy loss of $\Delta E = 20$ eV. What is the radius of the disc formed by a point object? How does the radius compare to atomic distances in a molecule?

26.3 Derive the Scherzer focus maximizing transparency of the CFT assuming an optimal value for the phase shift χ of 120°.

26.4 Derive the Scherzer expression for the point resolution in an electron microscope

$$d_0 = 0.65 \cdot C_s^{1/4} \cdot \lambda^{3/4}$$

26.5 What is the Scherzer resolution for a TEM operating at U_{acc} = 300 kV and C_s = 0.5 mm?

26.6 At what tilt angles θ are the path lengths s of electrons through a sample of thickness d 10% longer, double, and triple than at θ = 0 ?

26.7 Compare advantages, disadvantages, and limitations of crystallography, NMR, and cryo-EM in structural biology.

REFERENCES

Bai, X. C., Fernandez, I. S., McMullan, G. and Scheres, S. H. (2013) Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2**: e00461.
· ribosome structures to near-atomic resolution by cryo-EM

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S. L., …, Subramaniam, S. (2016) 2.3 A resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science* **351**(6275): 871–875.
· EM structure of the 540 kDa ATPase p97, an anti-cancer drug target

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L. and Subramaniam, S. (2015) 2.2 A resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. *Science* **348**(6239): 1147–1151.
· cryo-EM structure of a comparatively small protein, the tetrameric 465 kDa β-galactosidase

Fischer, N., Neumann, P., Konevega, A. L., Bock, L. V., Ficner, R., Rodnina, M. V. and Stark, H. (2015) Structure of the *E. coli* ribosome-EF-Tu complex at < 3 Å resolution by Cs-corrected cryo-EM. *Nature* **520**(7548): 567–570.
· EM structure of the ribosome/EF-Tu complex

Franken, L. E., Grunewald, K., Boekema, E. J. and Stuart, M. C. A. (2020) A technical introduction to transmission electron microscopy for soft-matter: Imaging, possibilities, choices, and technical developments. *Small* **16**(14): e1906198.
· review article on cryo-EM.

Lu, P., Bai, X. C., Ma, D., Xie, T., Yan, C., …, Scheres, S. H. (2014) Three-dimensional structure of human γ-secretase. *Nature* **512**(7513): 166–170.
· EM structure of human γ-secretase

Orlova, E. V. and Saibil, H. R. (2011) Structural analysis of macromolecular assemblies by electron microscopy. *Chem Rev* **111**: 7710–7748.
· review article on cryo-EM with image sequences showing radiation damage, reaction of Thon rings as a function of defocus and astigmatism, and the effects of missing wedges in image reconstruction.

Scheres, S. H. (2014) Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife* **3**: e03665.
· describes how motion correction due to the electron beam can be corrected to yield higher resolution data

Singh, K., Bunzel, G., Graf, B., Yip, K. M., Neumann-Schaal, M., Stark, H. and Chari, A. (2023) Reconstruction of a fatty acid synthesis cycle from acyl carrier protein and cofactor structural snapshots. *Cell* **186**(23): 5054–5067.
· Several structures of the 2.6 MDa enzyme have been determined, some to better than 0.2 nm resolution.

27

Scanning probe microscopy and force measurements

The covalent and non-covalent interactions of molecules are associated with forces, calculated as the derivative of the energy function (Chapter 17, Sections 20.1.1 and 20.1.2). Force methods, as the name implies, apply calibrated forces to biomolecules or measure forces associated with biological processes. In scanning probe microscopy, forces between a small tip and the sample are used for imaging, revealing surface topology or surface elasticity (Section 27.1). Force spectroscopy is the collective term for single-molecule techniques that measure forces within or between molecules to probe the strength of interactions, or the structure, stability, and folding pathway of macromolecules. In this type of experiment, the force is measured as a function of the extension of the molecule, either by *atomic force microscopy* (Section 27.1), *optical* tweezers (Section 27.2.3), or *magnetic tweezers* (Section 27.2.4).

Forces are central in any kind of motion. For example, motor proteins generate forces when they move along their track, such as myosin that moves along actin filaments during muscle contraction, or kinesin that moves along tubulin filaments to transport its cargo through the cell. DNA and RNA polymerases exert forces on their DNA template during replication and transcription. Ribosomes exert forces on the translated mRNA. In force experiments, these forces can be measured, and the mechanochemical coupling of enzymes and motor proteins can be interrogated. While AFM and optical tweezers stretch molecules to generate force-extension curves, magnetic tweezers allow to apply torque to DNA to study DNA twisting, untwisting, and supercoiling.

27.1 SCANNING PROBE MICROSCOPY: SCANNING TUNNELING, SCANNING FORCE, AND ATOMIC FORCE MICROSCOPY

In scanning probe microscopy, surfaces are scanned by a small probe, and distance-dependent interactions of the probe with the surface are measured. *Scanning tunneling microscopy* (STM) measures the tunnel current (0.01–10 pA) between the probe and the surface. In *scanning force* or *atomic force microscopy* (SFM/AFM), a small tip is scanned across the surface, and the sum of all attractive and repulsive forces between the tip and the surface is measured (Figure 27.1). These forces are in the

range of several pN to 100 nN. Biological samples are spread on mica, whereas hydrophobic organic molecules are analyzed on graphite surfaces.

AFM provides information on surface topology with nanometer resolution, depending on the dimensions of the tip. The pointed tip is attached to a cantilever. Attractive or repulsive forces between the tip and the sample surface lead to deflection of the cantilever, which is measured as a function of position, yielding an image of the surface in terms of height (Figure 27.1). AFM measures the sum of several forces with different range, such as van der Waals forces, electrostatic forces, or capillary forces. By functionalization of the tip, specific forces can be measured, including conductivity, hydrophobicity, or binding forces between interacting molecules.

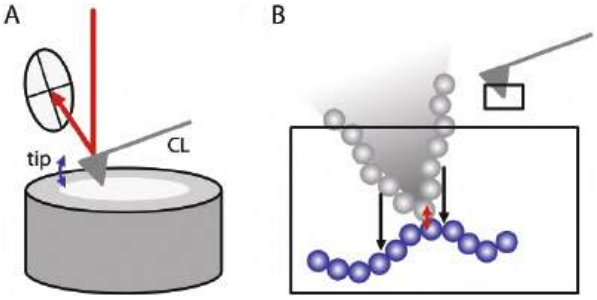


FIGURE 27.1: Principle of an AFM. A: A fine tip attached to a cantilever (CL) is scanned over the sample (light gray). Changes in sample height lead to a deflection of the cantilever (blue double-headed arrow). The deflection of the cantilever is detected by a laser (red) that is reflected from the cantilever to a position-sensitive detector. B: Schematic view of the tip (gray) and the surface (blue). The spheres show the tip and surface atoms. The force between tip and surface consists of attractive forces (black arrow) and repulsive forces (red arrow).

The extent of the deflection Δx of the AFM tip associated with a certain force F depends on the stiffness k of the cantilever, and follows Hooke's law (see Section 21.4.1, eq. 21.42):

$$F = -k\Delta x \qquad \text{eq. 27.1}$$

The lower the stiffness k, the higher the deflection associated with a certain force. Thus, cantilevers with lower stiffness allow measurements of smaller surface features, but are also more susceptible to perturbations that generate noise.

AFM can be performed in different modes. In the *contact mode* (Figure 27.2), the AFM tip is pulled directly across the surface. The deflection is measured either directly or the tip is kept at a constant distance by applying a feedback force. Such an experiment reveals the contours of the surface and the surface topology. For soft biological samples, the contact of the AFM tip with the surface can be destructive, and modes with intermittent or no contact between tip and sample are preferred. In the *tapping mode* (Figure 27.2), the AFM tip oscillates about 100–200 nm above the surface. Interactions with the surface lead to a damping of these oscillations. A feedback force keeps the tip at constant distance to maintain the oscillation frequency. The image obtained in the tapping mode contains information on surface topology (height), but also on its elasticity and hardness. In *frequency-modulated AFM* (non-contact mode, Figure 27.2), the AFM tip undergoes a low amplitude frequency-modulated oscillation (less than in the tapping mode) during surface scanning. The modulation is altered by (van der Waals) interactions with the surface. The same feedback system is used as for the tapping-mode experiment.

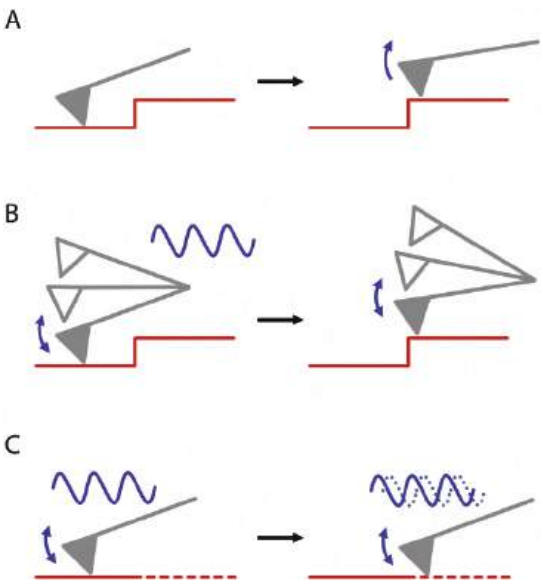


FIGURE 27.2: AFM modes and examples of AFM images. A: Contact mode. The tip (gray) is scanned over the surface (red) in direct contact. The topology of the surface leads to a deflection of the tip (right). B: Tapping mode and topology. The oscillating tip is moved across the surface, which minimizes the contact time. The step in the surface is detected as a deflection of the oscillating tip. C: Tapping mode and softness. The oscillating tip is scanned across the surface. The softer the surface, the larger the phase shift between the actual tip oscillation and the driving oscillation.

Recent developments in AFM methodology aim at imaging with increased, sub-second time-resolution (high-speed AFM) and at topological imaging in combination with fluorescence detection. In addition to imaging applications, AFM can also be used to measure inter- or intramolecular forces (see Section 27.2.2). Force spectroscopy by AFM is a powerful tool to analyze protein and nucleic acid folding and stability, forces exerted by enzymes on nucleic acids or filaments, and binding forces in general.

27.2 FORCE MEASUREMENTS

Force experiments are performed on single molecules, and therefore avoid limitations due to ensemble averaging (Section 25.7.2). They measure the mechanical response of single molecules without the need to synchronize and directly reveal the distribution of the observed parameter instead of the mean. Force experiments therefore reveal folding pathways, their probabilities, the existence of off-pathway and rare on-pathway intermediates, and kinetically trapped species. They detect movement and pausing of enzymes, their pausing probabilities, and the pause lengths. Before we discuss the different techniques to measure forces, we first have a look at the effect of applied forces on the energy landscape of molecules and on the kinetics of inter-conversion of states.

27.2.1 EFFECT OF FORCES ON THE ENERGETICS AND KINETICS OF MACROMOLECULES

Force is a state variable that affects equilibrium constants and rate constants, much as the state variables pressure p and temperature T. The free energy change dG now not only depends on temperature T and pressure p (see eq. 1.119 in Section 1.6.2), but also on the force F according to

$$\mathrm{d}G = -S\,\mathrm{d}T + V\mathrm{d}p - x\,\mathrm{d}F \qquad \text{eq. 27.2}$$

At constant pressure and temperature, eq. 27.2 simplifies to

$$dG = -x\,dF \qquad \text{eq. 27.3}$$

Integration from $F = 0$ (no force applied) to F (force F applied)

$$\int_{F=0}^{F} dG = -\int_{F=0}^{F} x\,dF \qquad \text{eq. 27.4}$$

yields

$$G(F) = G_0 - xF \qquad \text{eq. 27.5}$$

G_0 is the free energy in the absence of force. The linear term $x{\cdot}F$ leads to a tilting of energy profiles (Figure 27.3).

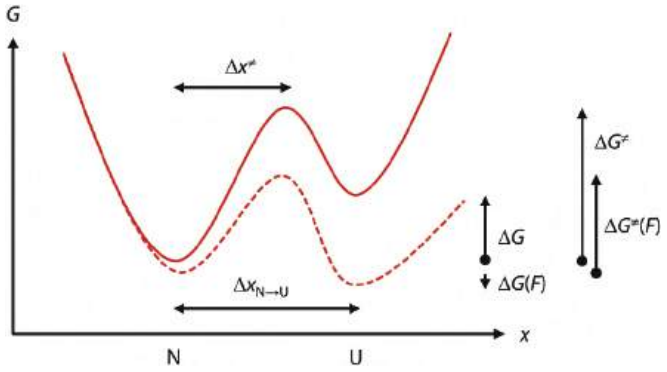


FIGURE 27.3: Tilting of energy landscapes by external forces. The energy landscape for unfolding of a protein from its native state N to its unfolded state U is tilted in the presence of an external force. High forces favor states with larger extension (eq. 27.5). The external force F therefore stabilizes the transition state of unfolding, and even more so the unfolded state with the highest extension. As a consequence, the energy barrier for unfolding, $\Delta G^{\neq}$, is decreased to $\Delta G^{\neq}(F)$, and unfolding is accelerated. The free energy change for unfolding is changed from ΔG to $\Delta G(F)$, and the equilibrium constant K_{unfold} increases.

From eq. 27.5, we see that the application of force stabilizes more extended species, and therefore changes equilibrium constants between species with different dimensions. For the force dependence of changes in free energy $\Delta G(F)$, we obtain

$$\Delta G(F) = \Delta G_0 - F\Delta x \qquad \text{eq. 27.6}$$

where Δx is the change in extension from reactant to product. In accordance with the principle of Le Chatelier (Section 2.1.2), the application of force thus favors reactions with an increase in extension, *i.e.* unfolding reactions. In contrast to p and T, force is a vector. The effect of force on the free energy ΔG thus depends on the direction of force application.

Forces also affect rate constants in reactions where the extension of the molecule is changed in the transition state, such as unfolding reactions:

$$k(F) = k_0 \exp\left(\frac{F dx^{\neq}}{k_B T}\right) \qquad \text{eq. 27.7}$$

k_B is the Boltzmann constant, $dx^{\neq}$ is the change in extension in the transition state, and k_0 is the unfolding rate constant at zero force. Note that unfolding pathways in mechanically induced unfolding transitions are most likely different from the pathways in thermal or solvent-induced unfolding. The rate constant k_0 therefore may or may not be comparable to unfolding rates from ensemble experiments. Rate constants for folding and unfolding can be determined in force experiments at constant force, where the molecule dynamically changes between the folded state (compact, low extension) and the unfolded state (high extension). From the distribution of dwell times in the folded and unfolded state, rate constants of unfolding and folding can be determined (Section 25.7.4). Alternatively, unfolding can be performed at different pulling rates, and rate constants are determined from the distribution of unfolding forces.

The dynamic range of force spectroscopy covers extensions as small as 0.3 nm corresponding to movements on DNA by a single bp up to 100 µm when entire living cells are manipulated. Forces range from fractions of a pN applied in unfolding of RNA molecules up to the high pN to nN range, which is sufficient to break chemical bonds. In AFM-based force spectroscopy, forces range from a few pN up to nN and extensions between 0.5 and 10^4 nm can be measured. With optical and magnetic tweezers lower forces can be measured or applied (0.1–100 pN for optical tweezers, 0.001–100 pN for magnetic tweezers). Magnetic tweezers reach about 5 nm resolution; with optical tweezers, distance changes down to 0.1 nm can be resolved.

27.2.2 Force spectroscopy by AFM

We have discussed before how we can obtain images of a surface by atomic force microscopy (AFM), either in terms of height or elasticity/softness with sub-nanometer resolution (Section 27.1). For force spectroscopy applications of AFM, the cantilever is not scanned laterally over the sample as in the imaging mode, but moved in the vertical direction. The molecule of interest is attached to a surface, and its free end is coupled to the cantilever (Figure 27.4). It is often sufficient to attach the biomolecule of interest to the tip by non-specific adsorption. Alternatively, the molecule can be attached specifically to functionalized AFM tips that are commercially available. The cantilever functions as a handle: by moving the sample away from the cantilever, the surface-cantilever distance increases. The molecule is extended between AFM tip and surface as the pulling force increases. The force causes deflection of the cantilever, and the force magnitude can be calculated from the measured deflection.

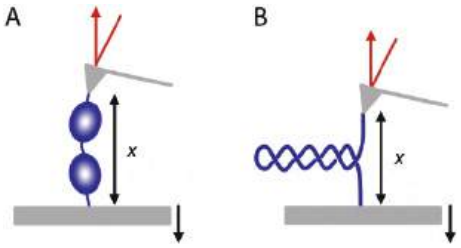


FIGURE 27.4: Force spectroscopy by AFM. A: Unfolding of a two-domain protein. The protein is attached to the surface on one end, and to the AFM cantilever on the other end. By moving the stage downward (black arrow), the distance between surface and cantilever is increased, and the molecule is stretched. The extension x of the molecule is the distance between surface and cantilever, the force is determined from the deflection of the cantilever (measured by laser light reflection, red). B: Unfolding of a DNA or RNA hairpin.

Before forces can be measured, the AFM needs to be calibrated. For small displacements, the cantilever deflection follows Hooke's law (eq. 27.1), meaning the force and the resulting displacement from its equilibrium position are proportional. By applying known forces and measuring the corresponding change in position Δx, the stiffness k of the cantilever can be determined. Once the stiffness is known, the deflection of the cantilever is recorded during force experiments along

with the distance of the cantilever from the surface. The deflection is then converted to forces using Hooke's law. Plotting the force as a function of the extension of the molecule, *i.e.* the distance between surface and tip, yields a *force-extension curve* (Figure 27.7). The forces associated with the stretching and unfolding of biomolecules are typically in the piconewton range, and changes in extension are on the nanometer scale.

If the biopolymer is rigid, the force required for its extension is high and roughly proportional to the extension. For elastic molecules, larger extensions are achieved with relatively small forces. The response of elastic polymers to applied forces can be described by different polymer models, such as the *Gaussian chain* (GC), *freely jointed chain* (FJC), and *worm-like chain* (WLC) models (Box 27.1). These models provide relationships between force and extension for elastic polymers that can be used to describe experimentally obtained force-extension curves and to extract information on the mechanical properties of the molecule of interest.

BOX 27.1: EXTENSION OF ELASTIC POLYMERS BY FORCE: GAUSSIAN CHAIN, FREELY JOINTED CHAIN, AND WORM-LIKE CHAIN.

There are two limiting cases for the behavior of a linear polymer: it can behave as a rigid rod or it can act as a completely flexible chain (Figure 27.5). The *Gaussian chain* model describes the conformation of a polymer by a Gaussian distribution of the end to end vector $\boldsymbol{x}$. For small end-to-end distances x compared to the contour length L, the force to stretch the polymer is a linear function of the stretching distance x:

$$F = 3\frac{k_B T}{L \cdot l_K} \cdot x \qquad \text{eq. 27.8}$$

k_B is the Boltzmann constant, and l_K is the Kuhn length (see below). Thus, a Gaussian chain behaves like an ideal spring and follows Hooke's law (eq. 27.1) for small forces. For larger forces and an increasing alignment and extension of the molecule, however, eq. 27.8 does not hold, and the polymer model needs to be refined. The *freely-jointed chain* (FJC) model describes the polymer as a chain of N rigid elements of length l_K, the Kuhn length. The *contour length* L of the polymer, *i.e.* the length along its backbone, is N times l_K. From the FJC model, the following relationship between force and extension of the polymer can be derived (eq. 27.9).

$$x = N \cdot l_K \left(\coth\left(\frac{F \cdot l_K}{k_B T} \right) - \frac{k_B T}{F \cdot l_K} \right) \qquad \text{eq. 27.9}$$

Eq. 27.9 can be used to describe measured force-extension curves to extract l_K and N. For small forces ($F \ll k_B T/l_K$), eq. 27.9 can be approximated and rearranged to

$$F = 3\frac{k_B T}{N \cdot l_K^2} \cdot x \qquad \text{eq. 27.10}$$

which is identical to eq. 27.8 and corresponds to the behavior of a Gaussian chain following Hooke's law (eq. 27.1). For large forces ($F \gg k_B T/l_K$), eq. 27.9 gives

$$F = \frac{k_B T}{l_K} \cdot \frac{1}{1 - \frac{x}{L}} \qquad \text{eq. 27.11}$$

Eq. 27.11 diverges for $x \to L$, *i.e.* when the molecule approaches its contour length. The behavior of polysaccharides under force is often well described by the FJC model.

(*Continued*)

The *worm-like chain* (WLC) model assumes that the polymer is flexible along the chain, with an irregular, position-dependent curvature. It covers all possible cases between rigid rods and fully flexible chains. The polymer is described by its contour length L and the *persistence length* l_p, a measure of the polymer stiffness. The persistence length describes over which distance the correlation in direction of the tangents to the curve is lost, or on which length scale the polymer appears linear. Below the persistence length, the polymer behaves like a rigid rod. Expressions for the force F as a function of the extension x have been derived from the WLC model using numerical procedures. The dependence of force on extension is described by

$$F(x) = \frac{k_B T}{l_p} \cdot \left(\frac{1}{4\left(1 - \frac{x}{L}\right)^2} + \frac{x}{L} - \frac{1}{4} \right) \qquad \text{eq. 27.12}$$

Eq. 27.12 can be used to describe the measured force-extension curves for biopolymers (Figure 27.5). The WLC model describes stretching of single and double-stranded DNA and RNA, and of polypeptides without structure. The persistence length of double-stranded DNA is 50 nm, for RNA it is 62 nm. Single-stranded nucleic acids have a much smaller persistence length of ~1 nm.

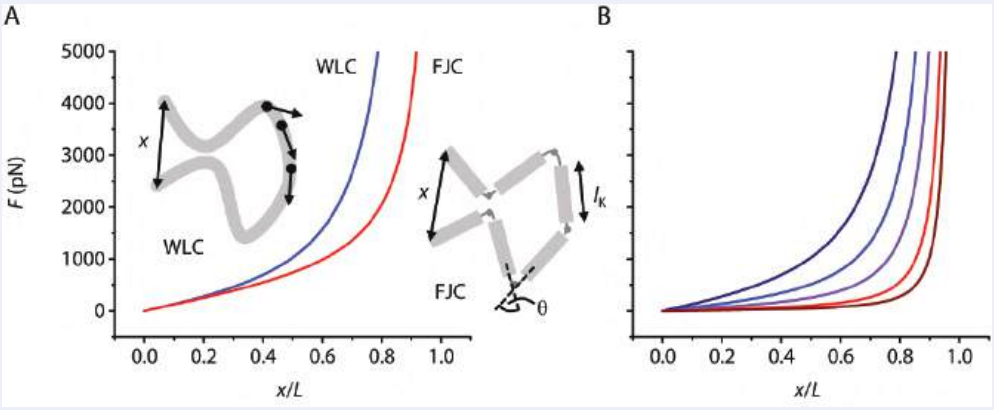


FIGURE 27.5: Force-extension curves according to FJC and WLC models. A: Comparison of force-extension curves according to the WLC and FJC models. Contour length L = 3400 nm (100 bp), persistence length l_p = 50 nm, T = 298 K, Kuhn length l_k = 100 nm. B: Force-extension curves according to the WLC model with different persistence lengths. Dark blue: l_p = 50 nm, blue: l_p = 100 nm, purple: l_p = 200 nm, red: l_p = 500 nm, brown: l_p = 1000 nm.

In the presence of structural changes, the force-extension curve is not continuous as described by the different polymer models (eq. 27.9, eq. 27.12), but contains jumps representing structural transitions. Single- and double-stranded DNA, for example, show different elasticities: double-stranded DNA is more rigid than single-stranded DNA. In other words, the persistence length for double-stranded DNA is larger than for single-stranded DNA. The persistence length of double-stranded DNA is 50 nm, whereas single-stranded DNA has a much smaller persistence length of less than 1 nm. The WLC model thus predicts different force-extension curves for stretching of single- and double-stranded DNA (Figure 27.6). Transitions between double- and single-stranded DNA, such as the dissociation of double-stranded DNA, appear as jumps in the force-extension curve, from one WLC curve to the other.

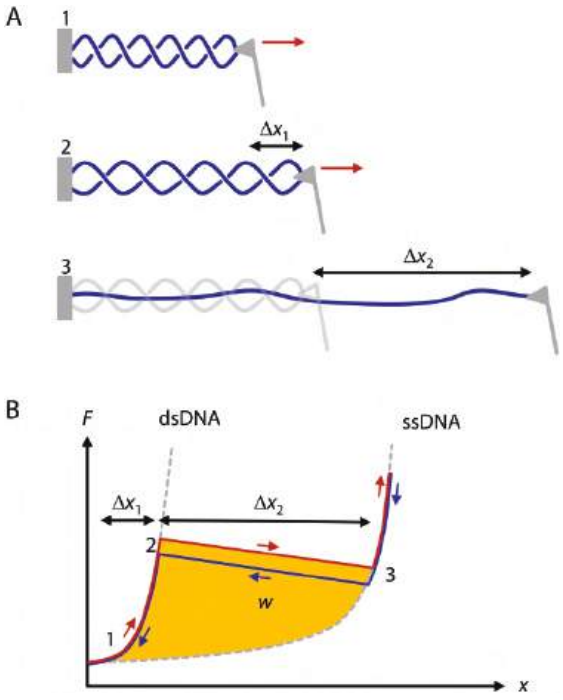


FIGURE 27.6: Force-extension curve for double-stranded DNA. A: To measure force-extension curves, double-stranded DNA is attached to a surface on one end. The other end is attached to the cantilever (1). When the cantilever is moved, the DNA is stretched (1→2), and its extension increases by Δx_1. At a certain force (2), the DNA strands dissociate (2→3), accompanied by a sudden increase in extension (Δx_2). B: Force-extension curve for stretching (red) and release (blue). During stretching of the double-stranded DNA (1→2), the force-extension curve can be described by the WLC model (eq. 27.12) with the persistence length and contour length of the double-stranded DNA as parameters. When the DNA strands dissociate (2→3), a sudden increase in extension is observed. After that (3), single-stranded DNA is extended, and the force-extension curve follows a WLC model with the persistence length and contour length of the single-stranded DNA. The area under the force-extension curve (orange) reflects the work *w* performed during strand separation. For reversible transitions, it equals the free energy change ΔG. Reversibility can be tested by measuring stretch-release curves. For reversible transitions, both curves superimpose. For irreversibility, a hysteresis is observed: unfolding upon stretching (red) occurs at higher forces than refolding during release (blue).

A force-extension curve directly provides information on the change in end-to-end distance associated with the underlying structural change, which is related to molecular dimensions. From the areas under the force-extension curves, the work *w* performed during the structural transition can be calculated by integrating the force (in pN) over the distance (in nm):

$$\int F \mathrm{d}x = w \qquad \text{eq. 27.13}$$

The values are typically given in pN·nm, which corresponds to 10^{-21} J or 1 zJ:

$$1\,pN \cdot 1\,nm = 10^{-12}\,N \cdot 10^{-9}\,m = 10^{-21}\,J = 1zJ \qquad \text{eq. 27.14}$$

1 pN·nm is on the order of the thermal energy of molecules (0.24 $k_B T$ at 298 K).

For molecules that consist of multiple independently (un)folding domains, force-extension curves contain multiple transitions reflecting the unfolding of each domain (Figure 27.7). In these cases, AFM experiments provide information on the underlying change in extension and the work performed during each structural transition.

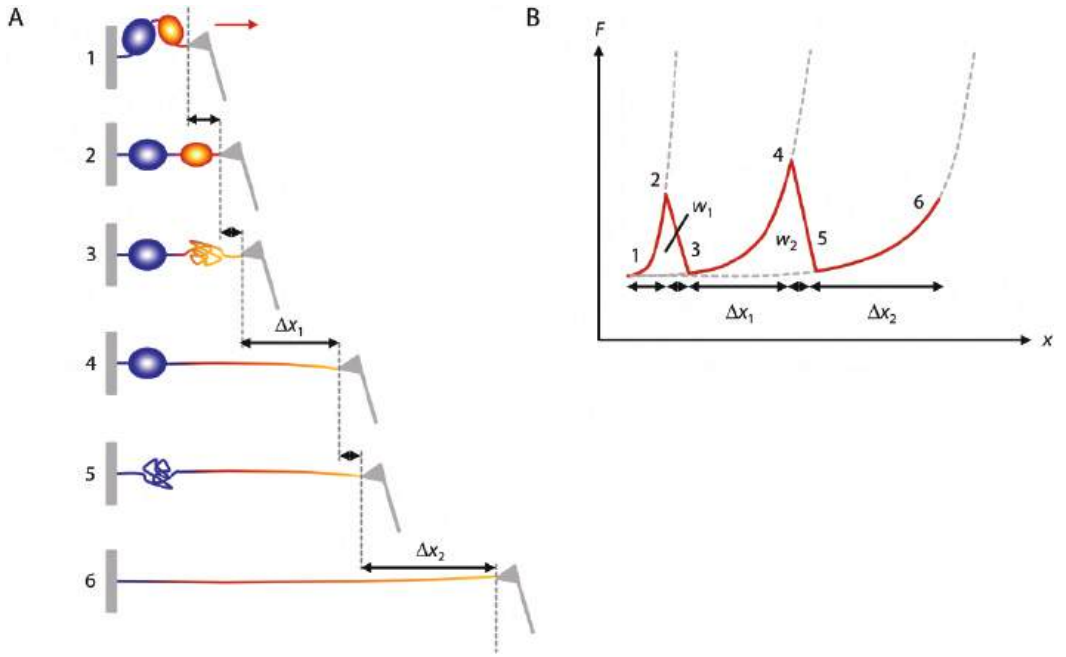


FIGURE 27.7: Measuring force-extension curves by AFM. A: Unfolding of a two-domain protein by AFM. The protein is immobilized on a surface (gray) with one end, and attached to a cantilever (gray triangle) with the other end. By moving the cantilever, the molecule is extended and unfolded by force. At the beginning (1→2), the coiled termini are stretched, and little force is required to extend the protein. After that, increasing force achieves little change in end-to-end distance. The energy is required to overcome interactions within the folded protein domains. The orange domain unfolds first. Once the interactions have been overcome (2), the force drops (2→3). While the polypeptide chain of the orange domain is extended (3→4), increasing the end-to-end distance requires little force. Further pulling then unfolds the blue domain (4→5). After that, small forces again achieve large changes in end-to-end distance until the polypeptide chain is stretched to complete extension (6). The area under the curve corresponds to the work performed during unfolding of each domain (w_1, w_2), and equals ΔG for reversible transitions.

For reversible transitions, this work is the associated change in free energy ΔG. Reversibility can be tested by measuring *stretch-release curves* (Figure 27.6), where the molecule is unfolded by increasing force (stretch), and then allowed to refold while the force is decreasing (release). If both curves superimpose, the transition is reversible. A hysteresis indicates lack of reversibility.

AFM has been used to study the mechanical properties of nucleic acids, polypeptides, and polysaccharides and to investigate the unfolding of proteins and DNA. AFM is also suitable to measure unfolding of membrane proteins in the context of a biological membrane (Figure 27.8). The membrane is deposited on the surface, and the AFM tip is moved close to the surface and attaches to the protein of interest. Pulling the cantilever upward then leads to unfolding of the attached protein until the polypeptide chain is removed completely from its membrane environment. During this process, force-extension curves for unfolding can be measured. AFM unfolding studies of bacteriorhodopsin and halorhodopsin, the water channel aquaporin, and a bacterial Na^+/H^+ antiporter have shown that membrane proteins fold and unfold differently from soluble proteins. While soluble proteins typically show cooperative transitions, membrane proteins can be unfolded one element after the other. By attaching the AFM tip to the terminus of the membrane protein, the first α-helix can be unfolded and extracted from the membrane, followed by the second, third, and so on. Reducing the applied force after unfolding of one or two helices allows re-formation of the helices and re-insertion into the membrane: the molecule of interest refolds. The force-extension curve for a seven-helix transmembrane protein such as bacteriorhodopsin therefore consists of multiple transitions. From these curves, the energy involved in the underlying structural transition and the length change per transition can be determined. By comparing different constructs lacking individual structural elements of the molecule of interest, observed transitions can be assigned to unfolding of individual elements of the molecule.

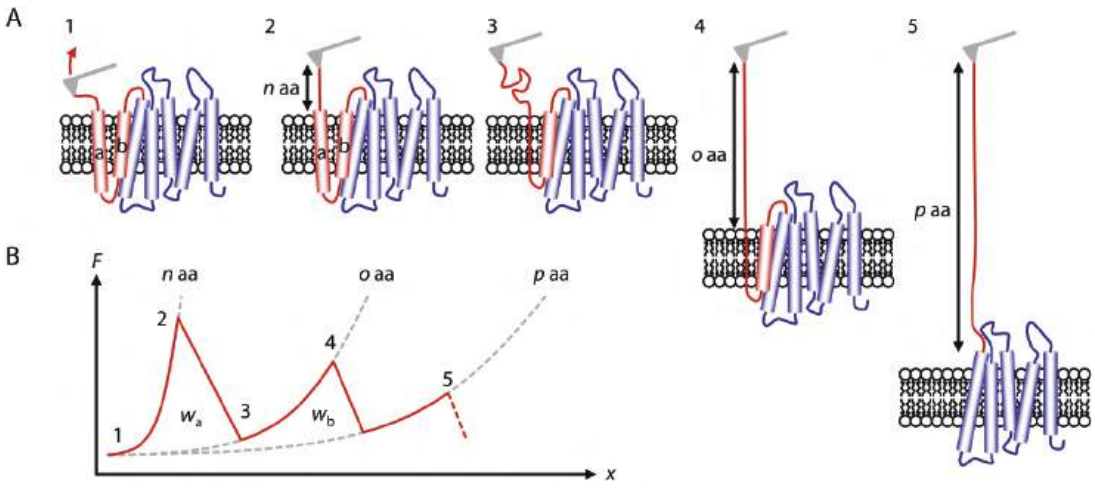


FIGURE 27.8: Force-induced unfolding of membrane proteins. A: Application of force to membrane proteins unfolds one structural element after the other (here: helices a and b). First, the terminus is extended (1→2), then helix a becomes unfolded (2→3) and fully extended (3→4), followed by unfolding of helix b and its full extension (4→5). B: Force-extension curve. The broken lines indicate the force-extension curves according to the WLC model. *n*, *o*, and *p* are the numbers of amino acids (aa) of the unfolded chain.

Force spectroscopy by AFM also allows measuring rupture forces to determine binding energies and equilibrium dissociation constants (Figure 27.9). In such an experiment, a ligand is covalently coupled to a functionalized AFM tip, and brought close to its surface-attached binding partner to allow complex formation. By pulling the cantilever away from the surface (or strictly speaking by moving the surface away from the cantilever), the force on the complex increases until the external force exceeds the mechanical stability of the complex. At this *rupture force*, the ligand dissociates from its binding partner, and the force drops to zero. As the method is not an equilibrium approach, the work performed, *i.e.* the integral under the force-extension curve, does not correspond to ΔG of the dissociation reaction. However, ΔG can be calculated from the distribution of rupture forces and the distribution of work values determined over many experiments (eq. 27.15):

$$\Delta G = -k_B T \cdot \ln\left\langle \exp\left(-\frac{w}{k_B T}\right)\right\rangle \qquad \text{eq. 27.15}$$

k_B is the Boltzmann constant.

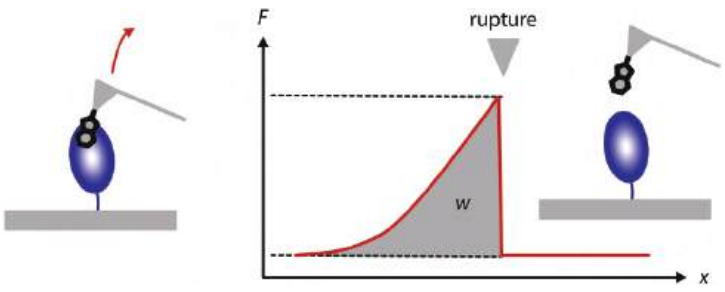


FIGURE 27.9: Measuring rupture forces by AFM. A ligand is attached to the AFM tip, and the tip is moved close to the surface to allow binding of the ligand to its binding partner. Retraction of the cantilever leads to disruption of the interaction, associated with a certain rupture force and work. As the process is irreversible, the work performed is not equal to the free energy change ΔG. The change in free energy and thus the stability of the interaction can be calculated from the distribution of the work for many rupture events (eq. 27.15).

Novel developments in AFM force spectroscopy aim at combining different types of information, such as AFM imaging with force mapping and force spectroscopy at individual points, or a combination of AFM with fluorescence detection on isolated molecules and in living cells.

27.2.3 Optical tweezers

In force experiments with optical tweezers, the molecule of interest is attached with one end to a polystyrene or silica bead that is trapped in a laser focus. The second end of the molecule is attached to a second bead that is held and moved to exert a force. Before we discuss how these experiments are performed and what we can learn from them, we will first introduce the concept of optical trapping. Optical tweezers are based on trapping of micrometer-sized dielectric objects in a laser focus. The trapping effect is caused by the gradient in light intensity, from high intensity in the focus to low and finally no intensity further away from the focus. The incident laser light is refracted by the object, and induces fluctuating dipoles in the object. These dipoles, as any dipole in an inhomogeneous electric field, experience a force in the direction of the field gradient. This *dipole force* is proportional to the polarizability of the object and to the intensity of the gradient. With tightly focused laser light, strong field gradients are generated, leading to large forces. If the object is moved out of the focus, this force acts as a three-dimensional restoring force, pulling the object back toward the focus (Figure 27.10). Scattering and reflection of light by the object, however, lead to forces that cause a movement of the object in the propagation direction of the light, and thus outside of the focus. If the gradient force due to refraction overcompensates the forces due to scattering, the object is stably trapped in the focus (Figure 27.10).

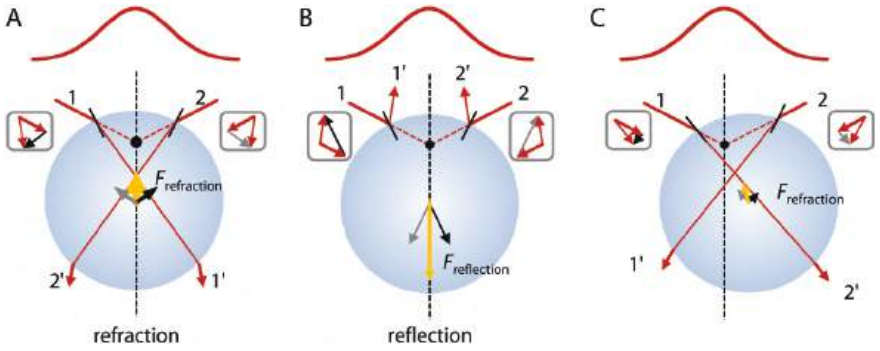


FIGURE 27.10: Principle of optical trapping. A polystyrene bead (blue) in the light path refracts and reflects the incident light. For clarity, only the two outer rays of the focused light bundle are depicted (1, 2). The red Gaussian distribution indicates the intensity gradient around the optical axis (black broken line). A: Refraction leads to a change in direction of the incident light (1, 2) after exiting the polystyrene bead (1′, 2′). The change in direction (black for beam 1, gray for beam 2, insets) and thus the impulse of the light leads to an opposite impulse change of the bead, and to a resulting force (yellow) on the bead. The restoring force is always directed toward the light focus (black circle) and leads to a trapping of the bead in the focus. B: Reflection of the incident light at the surface of the bead also changes the propagation direction of the light (inset). The opposite force on the bead now pulls it out of the focus and counteracts the trapping force. C: Displacement of the bead off the optical axis leads to a lateral force that traps the bead on the optical axis.

Typically, an infrared laser is used for trapping. The laser light is focused with a high numerical aperture objective. The rays at highest angles contribute most to the *trapping force*. The trapping force and the stability of the optical trap therefore increase with the numerical aperture of the objective (Figure 27.11). Trapping forces can be increased even further using dual-beam optical tweezers, where two opposing laser beams are focused in the same spot by two objectives (Figure 27.11). In this case, the scattering forces cancel, but the gradient forces are added, leading to more stable trapping. Optical trapping forces allow direct trapping of cells or organelles, as well as lipid vesicles, but are too small to trap biological molecules. Smaller molecules are attached to polystyrene or silica beads of about 1 µm in diameter, and these are trapped instead.

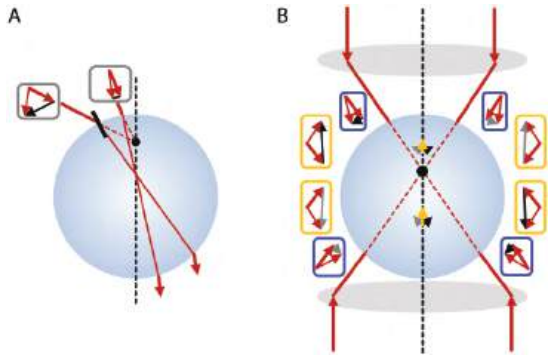


FIGURE 27.11: Increasing the trapping force. A: Outer rays contribute more to the trapping force than the inner rays of the light bundle. The higher the numerical aperture of the objective used, the higher the trapping force. The direction for an outer ray and an inner ray of the light bundle before and after the bead (red) and the change (black) are depicted in the inset. B: A dual-beam optical trap. If two opposite beams are focused into the same spot, the trapping forces acting on the bead add up, whereas the forces from reflection cancel. Both effects lead to more stable trapping. Orange inset: Changes in the propagation direction of the light due to reflection. Blue inset: Change in propagation direction due to refraction.

The optical trap holds the bead in its equilibrium position. However, external forces on the bead lead to small displacements Δx of the center of the bead from this equilibrium (Figure 27.12). Within short distances from the focus, the optical trap follows Hooke's law (eq. 27.1): displacements of the bead out of the focus depend linearly on the associated force. The ratio $F/\Delta x$ is the stiffness k of the trap. As discussed before for AFM measurements, the stiffness of optical tweezers has to be calibrated before experiments can be performed. This can be done through the hydrodynamic drag on the bead, by analyzing the thermal fluctuations of the bead, or by stretching a tethered DNA in a reference measurement. Once the stiffness is known, the displacement of the bead within the optical trap is measured and converted into the force using Hooke's law.

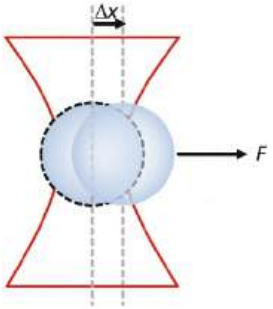


FIGURE 27.12: Response of a trapped bead to external forces. An external force F leads to a lateral displacement of the bead from its equilibrium position (black broken line) by Δx. When the stiffness k of the optical trap is calibrated, the force F acting on the bead can be calculated from the displacement Δx according to Hooke's law.

To perform a stretching experiment in an optical tweezers setup, the biomolecule of interest is attached on one end to the bead. On the other end it is attached to a second bead that is held in place by suction with a micropipette or by a second optical trap (Figure 27.13). The attachment of the biomolecule to the bead must be strong enough to withstand the forces applied during the experiment, and high-affinity interactions are necessary. DNA and also proteins are often labeled with biotin and coupled to streptavidin-coated beads (see also SPR in Section 29.3.4.2 and TIRF in Section 25.7.5; Figure 25.27). Another popular attachment method for DNA uses the interaction

between digoxygenin and the anti-digoxygenin antibody, and immobilizes digoxygenin-labeled DNA on beads functionalized with anti-digoxygenin antibodies. Proteins can also be immobilized *via* hexahistidine tags on Ni^{2+}-NTA-functionalized beads, although this interaction is less strong. Functionalized polystyrene beads are commercially available.

In optical tweezers experiments, the position of the bead relative to the trap center needs to be determined to calculate the force exerted on the bead. This is typically done by imaging the interference pattern formed by light from the trapping laser and the light scattered by the bead. As long as the bead is centered in the optical trap, the interference pattern is centrosymmetric. Small lateral changes in position lead to a displacement of the interference pattern and deviations from symmetry. The position is calibrated before the experiment by moving a bead on the surface of the sample cell by a known distance and measuring the positional signal. Interference pattern analysis allows detection of lateral movements of the bead by 0.1 nm on the millisecond time scale.

Optical tweezers can be operated in different modes. By applying *force ramps*, molecules attached to beads can be unfolded. In unfolding experiments, one end of the molecule is attached to a trapped bead, the other end to a second bead that is held in place with a micropipette or a second optical trap (Figure 27.13).

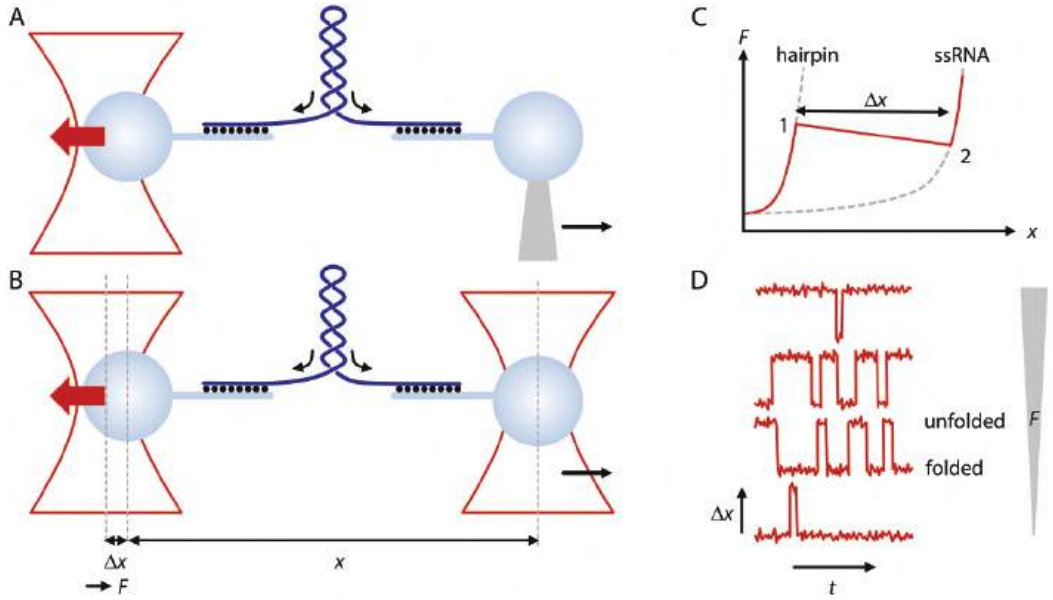


FIGURE 27.13: Optical tweezers experiments to monitor unfolding. The ends of the molecule of interest, here an RNA hairpin, are attached to polystyrene beads using DNA/RNA handles. One bead is in an optical trap, the second is either held by a micropipette (A) or by a second optical trap (B). Moving the micropipette or the second optical trap laterally (black arrow) exerts force on the molecule that pulls the trapped bead out of its equilibrium position. The force can be calculated from the displacement Δx. The extension x of the molecule is determined from the distance between the centers of the two beads. When the hairpin unfolds, the extension increases, and the force decreases. The trapped bead therefore moves to the left (red arrow). C: In force-ramp experiments, force-extension curves for unfolding of the hairpin (1→2) are measured. D: In force-clamp experiments, the force is maintained constant, and the extension of the molecule is measured as a function of time. The molecule switches back and forth between the folded state (low extension) and the unfolded state (high extension). Depending on the applied force, the molecule spends more time in the folded state (low force) or in the unfolded state (high force).

While proteins and DNA are attached directly to the beads, RNA molecules are often attached through DNA/RNA handles. To stretch the molecule of interest, the micropipette or the second optical trap is moved away. The force exerted is then determined from the displacement of the trapped bead from its equilibrium position. From the distance of the bead centers (minus the radii of the beads), the extension is calculated. Variations in bead radius lead to uncertainties for absolute

extensions, but extension changes can be measured down to < 1 nm. The force is continuously increased or decreased to induce unfolding of the biomolecule, and force-extension curves are measured. A structural transition of the molecule is associated with an abrupt change in force and extension: unfolding of a molecule leads to a sudden increase in its extension, with a concomitant drop in force, such that the trapped bead moves back toward the center of the optical trap. Conversely, a folding transition is measured as an abrupt decrease in extension and increase in force. For reversibly unfolding molecules, the work performed during the transition equals the underlying free energy change ΔG.

In the *force-clamp mode*, the position of the second bead is continuously adjusted during the experiment to maintain a constant force on the bead and the molecule of interest (force feedback). In this case, the displacement of the trap reflects changes in extension of the attached molecule. In force-clamp experiments, dynamic changes in the extension of the molecule attached to the beads can be monitored, *e.g.* due to dynamic transitions between folded (short) and unfolded (extended) states in equilibrium at constant force (Figure 27.13), due to shortening of single-stranded DNA that is converted into double-stranded DNA by polymerase activity, or DNA or RNA extension through action of a helicase. Using force-clamp experiments, the movement of RNA polymerase along DNA has been followed with single base pair resolution. Rate constants of structural transitions can be determined from the dwell time distribution of the molecule in each state at constant force. The equilibrium constant for folding is obtained from the ratio of the rate constants (see Section 8.1), and also allows the calculation of free energy changes. By performing force-clamp experiments at different forces, the force dependence of rate and equilibrium constants can be determined. Force-clamp experiments also allow determination of stalling forces, *i.e.* forces at which the net displacement is zero and translocation and activity cease. Stalling forces provide information on the force generated by these enzymes. The stalling force for RNA polymerase is ~25 pN, for DNA polymerase the stalling force is 34 pN. The stalling forces of the motor proteins myosin/actin and kinesin/microtubules are only around 5 pN.

Optical tweezers can also be operated in an *extension-clamp mode* where the extension is fixed by a feedback mechanism that constantly adjusts the position of the second bead. In this mode, structural transitions are measured as a change in force.

Force-jump experiments are used to measure unfolding and folding rate constants for the same molecule. First, the force is set to a high value at which the molecule unfolds. The time until unfolding occurs, *i.e.* the dwell time in the folded state, is measured. After that, the molecule is fully unfolded by further increasing the force. In the second part of the experiment, the force is then suddenly dropped to a value that allows refolding, and the dwell time of the unfolded state of the molecule is monitored. From dwell times, rate constants can be determined. In force-jump experiments, different structural transitions that occur at different forces can be studied separately. Each step is observed at a suitable force, as in force-clamp experiments, and rate constants are extracted from individual dwell times. These experiments allow to delineate multi-step folding pathways and the rate constants associated with each step, and to identify rate-limiting steps.

Optical tweezers have been used to study folding and unfolding of DNA and RNA. DNA and RNA hairpins unfold at forces of ~15 pN. For protein unfolding, AFM is often used because of the larger forces required in comparison to unfolding of nucleic acids.

Optical tweezers may also be used to measure forces on DNA or RNA exerted by processive enzymes such as polymerases, helicases, or ribosomes, or forces involved in movement of motors on their track, such as myosin moving along actin filaments, or kinesin moving along microtubules. Depending on the application and the type of information sought, different experimental configurations are used (Figure 27.14). In the *interaction assay*, one partner (for example, myosin) is attached to a bead in an optical trap, and the other partner (for myosin, an actin filament) is immobilized on the surface of the sample cell. The displacement of the bead in the optical trap reports on the relative movement of the partners to each other, and on the forces involved. This configuration is particularly suited to study movement of motor proteins along their track. In the *tethered assay* (Figure 27.14), an enzyme (for example, a polymerase moving along DNA) is attached to the optically

trapped bead. One end of the DNA is attached to the surface or to a second bead held by a micropipette. When the enzyme starts to move along the DNA, it either reels the DNA in or lets it out, depending on the direction of translocation. Translocation therefore leads to a displacement of the bead. At constant force, this configuration allows one to follow enzyme movement over long periods of activity. The position of the trapped bead directly reports on translocation of the enzyme. By measuring translocation rates at different forces, the mechanochemical coupling of the enzyme can be interrogated. The tethered assay is also used to study unfolding of nucleic acids. The *dumbbell configuration* (Figure 27.14) is a variation of the tethered assay in which the DNA end is attached to a second bead, which is held in its own, independent optical trap. During the enzymatic action on the DNA, one of the traps is continuously moved to keep the force constant. In this configuration less drift of the bead is observed than in a tethered assay.

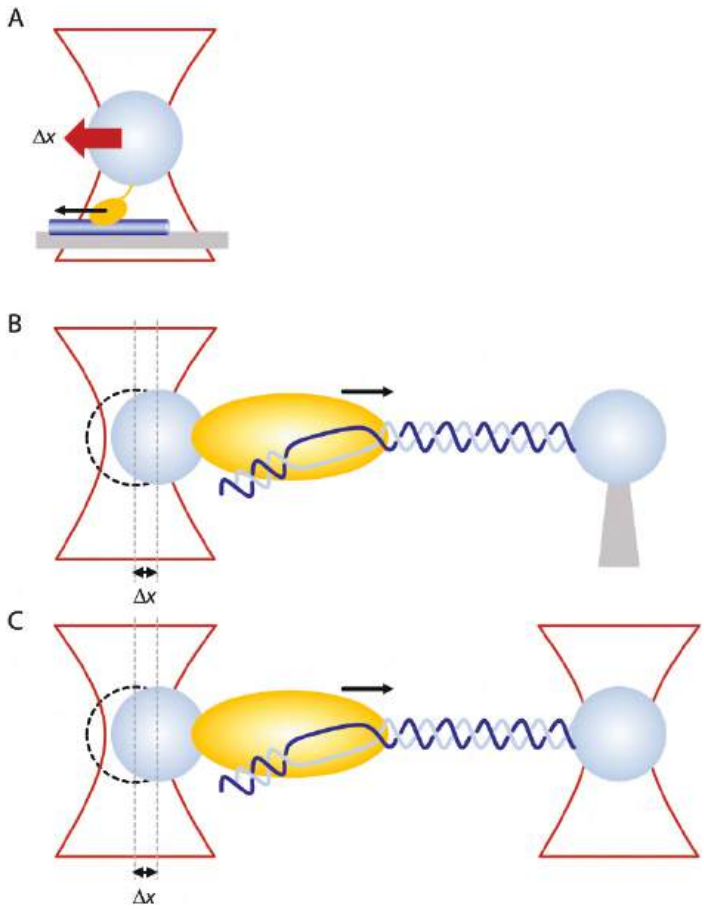


FIGURE 27.14: Different configurations of optical tweezers experiments. A: Interaction assay. A motor protein (orange) is immobilized on the polystyrene bead that is trapped. The motor protein moves along a track (blue) that is immobilized on the surface (gray). Movement of the motor along the track generates a force and causes a displacement of the trapped bead. B: Tethered assay. An enzyme (RNA polymerase, orange) is immobilized on the optically trapped bead. The DNA template (blue) is attached to a second bead that is held by a micropipette, the other end of the DNA is free. Movement of the RNA polymerase along the DNA leads to a displacement of the optically trapped bead. C: Dumbbell assay. In contrast to (B), the second bead is held by a second optical trap. Note that beads and molecules are not drawn to scale.

The dumbbell configuration offers the advantage of being able to move both optical traps within the sample chamber. Experiments can thereby be performed away from surfaces. Using microfluidic sample cells allows movement of the two optical traps and thus of the molecule of interest between different flow channels and for facile change of buffer conditions (Figure 27.15).

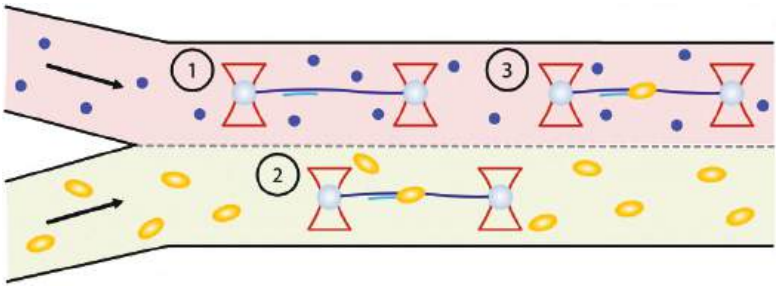


FIGURE 27.15: Optical tweezers experiments and microfluidic sample cells. A microfluidic sample cell with two flow channels contains deoxynucleotide triphosphates (blue spheres) in one channel (light red), and DNA polymerase (orange) in the other (light green). Beads (light blue) attached to the end of a single-stranded DNA (dark blue) with a hybridized primer (turquoise) are captured by optical traps (red, 1). By moving the optical traps, the DNA is moved into the second flow channel (light green, 2), where a polymerase molecule is allowed to bind to the DNA. The DNA-polymerase complex is then transferred back to the first flow channel where DNA synthesis starts (3).

Optical tweezers experiments have been used to study the effect of DNA or RNA binding proteins on DNA structure, such as wrapping of DNA by nucleosomes, to monitor movement of DNA and RNA polymerases, helicases, and ribosomes along DNA or RNA, and to investigate the mechanochemical cycles of motor proteins, such as actomyosin and kinesin, moving along microtubules.

The application spectrum of optical tweezers has been expanded by introducing torque and by combining optical tweezers experiments with fluorescence detection by TIRF, confocal or wide-field microscopy, or FRET. Optical trapping and force measurements are also possible *in vivo*. Multiplexing, *i.e.* observing multiple optical traps in parallel, allows manipulating more than one molecule at the same time (Box 27.2).

BOX 27.2: MULTIPLEXING WITH OPTICAL TWEEZERS: SIMULTANEOUS MANIPULATION OF TWO DNA MOLECULES.

Many proteins bridge DNA molecules by interacting with different regions of double-stranded DNA. The DNA in the bacterial nucleoid is associated with numerous proteins that mediate the compact folding of the DNA. The histone-like nucleoid structuring protein (H-NS) is a major determinant for the compaction and the structural organization of DNA in the nucleoid, and has also been ascribed a regulatory role for gene expression. H-NS binds to two DNA molecules and forms a bridge between them. The energy landscape of the DNA-DNA interaction mediated by H-NS has been probed with four optical traps that are used pairwise to trap two molecules of DNA. The quadruple trap (Q-trap) allows the application of force on bridged DNA molecules and to interrogate their mechanical stability (Figure 27.16). By moving the two bridged DNA molecules away from each other, the rupture of individual bridges was measured. Dame *et al.* (2006) Nature 444(7117): 387–390.

(*Continued*)

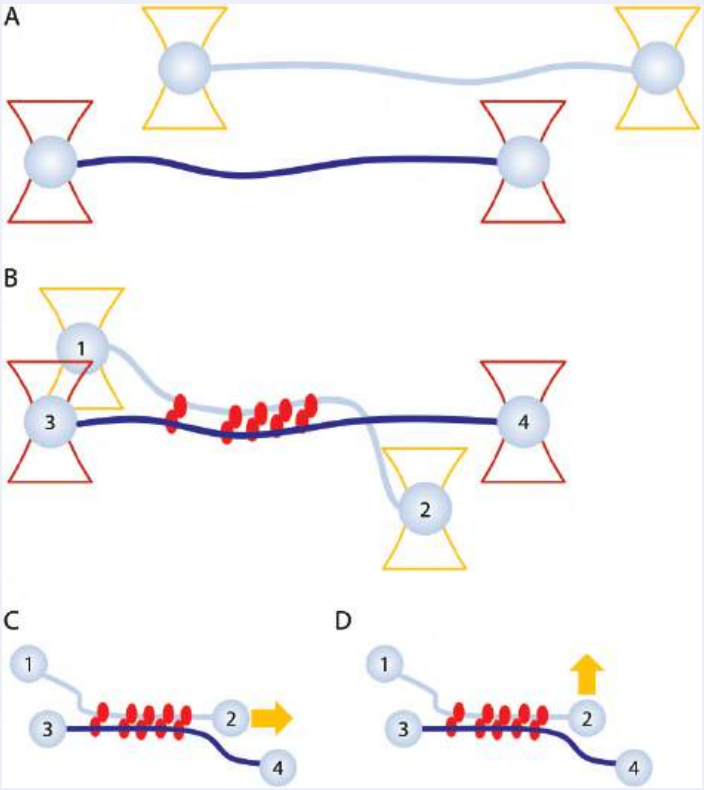


FIGURE 27.16: The quadruple trap and mechanical manipulation of two DNA molecules. A: With two pairwise optical traps (orange, red), two double-stranded DNA molecules (light blue, dark blue) can be manipulated. B: By moving the optical traps, the DNA molecules can be brought into close proximity, such that a bridging protein (red) can bind. C: By moving bead 2 to the right, shearing experiments can be performed. D: Moving bead 2 upward allows unzipping experiments in which stepwise disruption of individual bridges is observed.

27.2.4 Magnetic tweezers

Conceptually, magnetic tweezers work according to the same principle as optical tweezers, and also serve to apply forces to biomolecules or to monitor forces generated by biomolecules. In magnetic tweezers, magnetic beads are held in place by a pair of magnets. The magnetic field induces a magnetic moment in the bead. As a result, the bead experiences a force in the direction of the field gradient and proportional to the gradient. The field gradient is constant over the length scale of the movements of the magnetic bead, and no force feedback is required as in optical tweezers (see Section 27.2.3).

The molecule of interest is attached with one end to the magnetic bead, with the other end to the surface of the sample cell. Functionalized magnetic beads for the attachment of biomolecules are commercially available. By moving the magnets upward, away from the bead, the bead is pulled upward, and the attached molecule of interest is stretched as in optical tweezers stretching experiments. This type of experiment yields force-extension curves.

The strength of magnetic tweezers experiments is the capability to rotate the magnetic bead by rotation of the magnets. The torsion of the bead introduces twist into the attached molecule. By turning the magnetic bead, DNA can be over- or undertwisted. At a certain degree of twisting, further twist is converted into writhe (Section 19.5.3.2) and leads to "buckling" of the DNA and

plectoneme formation (Figure 27.17). Plectoneme formation leads to a strong decrease of the DNA extension. Twist and torque play an important role in all processes on DNA, including transcription, replication, and recombination.

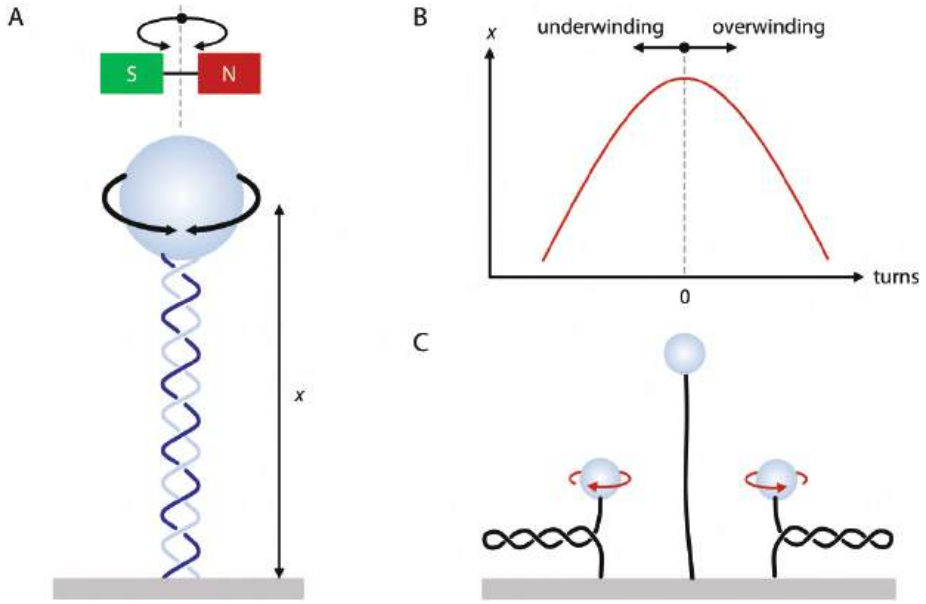


FIGURE 27.17: Twisting DNA by magnetic tweezers. A: A DNA molecule is attached with one end to the magnetic bead, with the other end to the surface. The magnetic bead is held by permanent magnets (red/green) at a height at which the DNA is extended. By turning the magnets, the DNA is under- or overwound. B: Extension of the DNA as a function of turns of the magnet. At low force, the DNA forms plectonemes when it is over- or underwound. C: Under torque, double-stranded DNA forms supercoiled regions, termed plectonemes.

The maximum force of a magnetic tweezers instrument is determined by the magnitude of the magnetic moment, which in turn depends on the size of the bead and is proportional to its volume. The second factor is the minimal distance of the magnets from the bead. The typical force range is 10–100 pN. The magnetic bead is the means to introduce torsional stress and, as for optical tweezers, is the object that is followed during the experiment by detecting the interference pattern of light scattered by the bead with unscattered light. Vertical movement of the bead moves the bead out of focus and changes the intensity distribution of the interference pattern. From changes in the number of concentric rings of the interference pattern, the height of the bead along the z-axis can be determined with about 10 nm accuracy. The height is calibrated before the experiment.

Magnetic tweezers have been used to probe the mechanical properties of DNA and RNA. They are also very useful to investigate the mechanisms of enzymes that act on DNA or RNA, such as DNA or RNA polymerases and helicases or DNA repair and replication proteins. Because of the facile introduction of twist and torque, magnetic tweezers are particularly suited to monitor the supercoiling or relaxation activity of DNA topoisomerases on DNA (Box 27.3) and to interrogate the mechanochemical coupling of topoisomerases. Magnetic tweezers have been used to study the mechanism of negative DNA supercoiling by gyrase and of positive supercoiling by reverse gyrase. They have also revealed the mechanism of topoisomerase I poisoning by the drug camptothecin.

BOX 27.3: MAGNETIC TWEEZERS, DNA TOPOLOGY AND DNA TOPOISOMERASES.

Using magnetic tweezers, negative or positive supercoils can be introduced into DNA by simple turning of the magnetic bead and thus under- or overtwisting of the DNA (Figure 27.18). Relaxation of these supercoils by DNA topoisomerases can then be monitored as an extension of the DNA. Measuring the extension as a function of time provides information on the step size of relaxation, and thus on the change in linking number per step. The time passing between two relaxation events provides kinetic information on DNA relaxation. Instead of generating plectonemes, magnetic tweezers can also be used to introduce crossings into DNA by twisting two double strands around each other (Figure 27.18). Using magnetic tweezers, it was shown that topoisomerase IV relaxes positive supercoils more rapidly than negative supercoils. Neuman *et al.* (2009) Proc. Natl. Acad. Sci. U S A 106(17): 6986–6991.

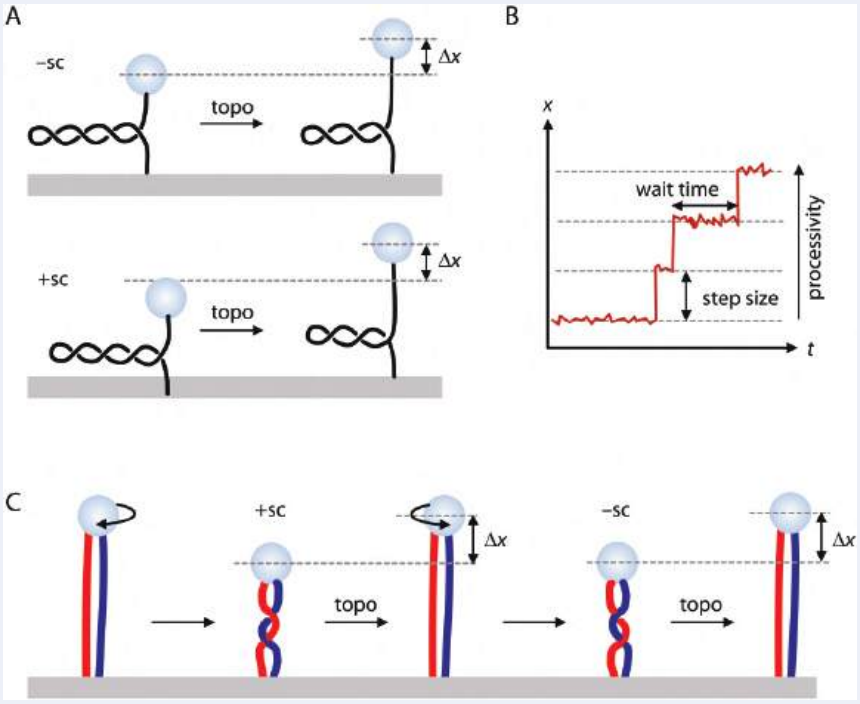


FIGURE 27.18: Monitoring relaxation of positive and negative DNA supercoils by topoisomerases. A: Under- and overtwisting of DNA generates plectonemes with negative and positive DNA supercoils. Relaxation of these supercoils leads to an increase in extension Δx of the DNA. B: Time dependent measurement of DNA extension during DNA relaxation. Schematic time trace of DNA extension as a function of time. Time traces provide information on the step size, the pausing time before a relaxation event, and the processivity of the DNA topoisomerase. C: Negative and positive crossings can also be introduced into DNA by twisting two double strands (red, blue) around each other. Again, relaxation of these supercoils by DNA topoisomerases (topo) is measured as an increase in DNA extension.

Although magnetic tweezers allow twisting of DNA molecules, they only provide a read-out for the extension of the molecule, not for its rotation. In a *rotor-bead assay* (Box 27.4), angular movement of the molecule of interest is monitored simultaneously with changes in extension. This method has been applied to follow topological changes in DNA during negative supercoiling by DNA gyrase and to interrogate mechanochemical coupling (Box 27.4).

BOX 27.4: ROTOR-BEAD ASSAY AND DNA SUPERCOILING.

A rotor-bead assay uses an additional bead that is attached to the DNA internally to provide a readout on twisting of DNA in optical or magnetic tweezers. The DNA molecule contains a nick that acts as a swivel. The rotor bead is attached below the nick. The DNA is held between magnetic tweezers at sufficiently high force to prevent buckling and plectoneme formation (Figure 27.19). To investigate the mechanochemical coupling of DNA gyrase, a binding site for gyrase was introduced in the bottom part of the DNA. When gyrase binds to the DNA and changes the linking number, the topological change can manifest as twist or writhe (Section 19.5.3.2). In the presence of force, writhing associated with a decrease in extension is disfavored, and the change in linking number leads to a change in twist of the DNA. This twist leads to a torque on the rotor bead. The angular position of the bead is a direct readout for the twist of the DNA below the nick, and its angular velocity reports on the torque. A change in linking number by 1 would result in a complete turn (360°) of the rotor bead, and a change in linking number by 2 is detected as a rotation by 720°.

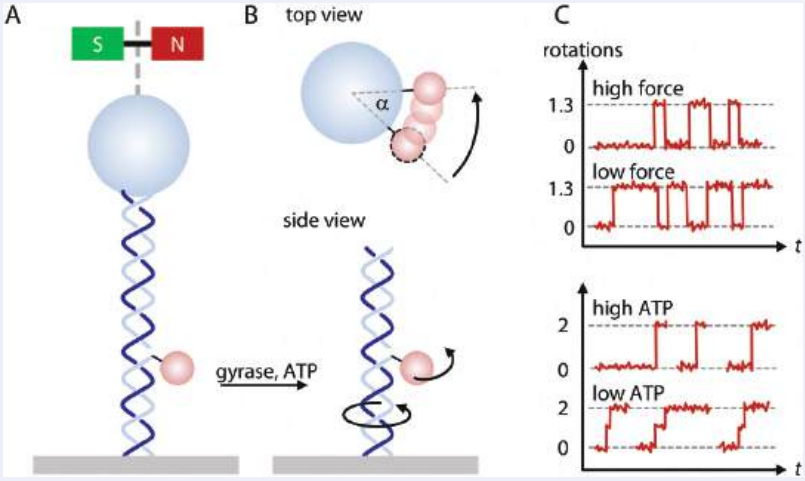


FIGURE 27.19: A rotor-bead assay to measure twist and torque. A: DNA with a nick is held under force in a magnetic tweezers setup. The DNA can swivel at the nick. A decrease in linking number of the DNA due to gyrase activity manifests itself as twist of the DNA below the nick, which exerts a torque on the rotor bead. B: Top view: The angle α reports on the twist, and the angular velocity of the rotor bead on the torque. The angle of the rotor bead is measured as a function of time. C: Schematic traces of the angle as a function of time to illustrate the behavior of DNA in the presence of gyrase at low and high force in the absence of ATP (top) and at low and high ATP concentrations (bottom). DNA wrapping and unwrapping in the absence of ATP is reflected in jumps between 0 and 1.3 rotations. At high ATP concentrations, each catalytic cycle leads to two complete turns of the rotor bead around the DNA. At low ATP concentration, pauses at one rotation are observed. Schematic traces for three different events are shown.

(*Continued*)

Using this procedure, it was confirmed that gyrase leads to two complete turns of the rotor bead during one catalytic cycle, in agreement with the linking number change by two. Gyrase binds DNA and wraps it around itself at the beginning of the catalytic cycle. In the absence of ATP, the rotor-bead angle alternates between two values, corresponding to DNA wrapped around gyrase (1.3 rotations) and DNA that is not wrapped (no rotation). When the force is increased, the time gyrase spends in the wrapped state is decreased. At high ATP concentrations, the rotor bead undergoes two complete rotations during each catalytic cycle. When the ATP concentration is reduced, a pause at about one rotation is detected. Under these conditions, progression through the catalytic cycle is reduced, and unwrapping efficiently competes with ATP binding and DNA supercoiling. Gore *et al.* (2006) Nature 439(7072): 100–104.

Recently developed magnetic tweezers make use of electromagnets instead of permanent magnets. The magnetic field of electromagnets can be changed by altering the current, which is much faster than moving conventional ferromagnets, and decreases the dead time. Recent developments also include multiplexing and the combination of magnetic tweezers with fluorescence detection and FRET.

QUESTIONS

27.1 An optical trap has a stiffness of $k = 0.05$ pN nm^{-1}. A polymerase pulls with $F = 22$ pN on the DNA attached to the bead in the optical trap. What is the displacement Δx? What would the displacement be in a magnetic tweezers setup ($k = 0.5$ pN nm^{-1}) or in an AFM experiment ($k = 100$ pN nm^{-1})?

27.2 The center of the polystyrene bead in an optical tweezers instrument can be determined with an accuracy of approx. 1 nm. A student wants to measure forces with a resolution of at least 5 pN. What is the stiffness the trap must have to achieve this?

27.3 Using optical tweezers at a constant force, you monitor the extension of a protein as a function of time for 25 s. During this time, the molecule spends a cumulated 5 s in the state with small end-to-end distance. The average time between two transitions that it spends in the extended state, with higher end-to-end distance, is 0.2 s. What are the rate constants for folding and unfolding, k_{fold} and k_{unfold}, and the equilibrium constant K_{fold}? What is the mean dwell time of the protein in the folded state?

27.4 In an optical tweezers experiment, an RNA hairpin spends ¼ of its time in the folded state at an applied force of 16 pN. This value increases to ¾ at 4 pN. What is the thermodynamic stability of this hairpin in the absence of force?

REFERENCES

Abbondanzieri, E. A., Greenleaf, W. J., Shaevitz, J. W., Landick, R. and Block, S. M. (2005) Direct observation of base-pair stepping by RNA polymerase. *Nature* **438**(7067): 460–465.
· investigation of the movement of RNA polymerase along DNA by force-clamp optical tweezers experiments with single base pair resolution

Abels, J. A., Moreno-Herrero, F., van der Heijden, T., Dekker, C. and Dekker, N. H. (2005) Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys J* **88**(4): 2737–2744.
· magnetic tweezers and atomic force study of the mechanical properties of double-stranded RNA

Ando, T. (2014) High-speed AFM imaging. *Curr Opin Struct Biol* **28**: 63–68.
· review article about instrumentation and applications in high-speed AFM

Arslan, S., Khafizov, R., Thomas, C. D., Chemla, Y. R. and Ha, T. (2015) Protein structure. Engineering of a superhelicase through conformational control. *Science* **348**(6232): 344–347.
· analysis of DNA unwinding processivity of the DNA helicase RepX using optical tweezers in combination with microfluidics

Ashkin, A. (1970) Acceleration and trapping of particles by radiation pressure. *Phys Rev Lett* **24**(4): 156–159.
· first description of the principle of optical trapping

Blehm, B. H., Schroer, T. A., Trybus, K. M., Chemla, Y. R. and Selvin, P. R. (2013) *In vivo* optical trapping indicates kinesin's stall force is reduced by dynein during intracellular transport. *Proc Natl Acad Sci U S A* **110**(9): 3381–3386.
· *in vivo* optical trapping and force measurements on kinesin and dynein

Bockelmann, U., Thomen, P., Essevaz-Roulet, B., Viasnoff, V. and Heslot, F. (2002) Unzipping DNA with optical tweezers: High sequence sensitivity and force flips. *Biophys J* **82**(3): 1537–1553.
· study on the mechanical stability of DNA hairpins

Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R. and Bustamante, C. (2003) Structural transitions and elasticity from torque measurements on DNA. *Nature* **424**(6946): 338–341.
· torque measurements on DNA using optical tweezers and a rotor-bead assay

Bustamante, C., Chemla, Y. R., Forde, N. R. and Izhaky, D. (2004) Mechanical processes in biochemistry. *Annu Rev Biochem* **73**: 705–748.
· review article on the effects of forces on folding, stability, and activity of proteins, highlighting motor proteins and enzymes as examples

Bustamante, C., Marko, J. F., Siggia, E. D. and Smith, S. (1994) Entropic elasticity of lambda-phage DNA. *Science* **265**(5178): 1599–1600.
· mechanochemical study of double-stranded DNA and description of the force-extension curve according to the worm-like chain model

Cisneros, D. A., Oesterhelt, D. and Muller, D. J. (2005) Probing origins of molecular interactions stabilizing the membrane proteins halorhodopsin and bacteriorhodopsin. *Structure* **13**(2): 235–242.
· comparative atomic force spectroscopy study of the mechanostability of halorhodopsin and bacteriorhodopsin

Comstock, M. J., Ha, T. and Chemla, Y. R. (2011) Ultrahigh-resolution optical trap with single-fluorophore sensitivity. *Nat Methods* **8**(4): 335–340.
· combination of optical tweezers with fluorescence detection at single-molecule sensitivity

Comstock, M. J., Whitley, K. D., Jia, H., Sokoloski, J., Lohman, T. M., Ha, T. and Chemla, Y. R. (2015) Protein structure. Direct observation of structure-function relationship in a nucleic acid-processing enzyme. *Science* **348**(6232): 352–354.
· a combined optical tweezers/confocal fluorescence microscopy study on the DNA helicase UvrD

Dame, R. T., Noom, M. C. and Wuite, G. J. (2006) Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature* **444**(7117): 387–390.
· analysis of the energy landscape of the DNA-DNA interaction mediated by H-NS by pairwise trapping of two molecules of DNA in a quadruple optical trap (Q-trap)

Dumont, S., Cheng, W., Serebrov, V., Beran, R. K., Tinoco, I., Jr., Pyle, A. M. and Bustamante, C. (2006) RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* **439**(7072): 105–108.
· analysis of translocation of hepatitis C NS3 helicase on RNA by optical tweezers with two base pair resolution

Francius, G., Alsteens, D., Dupres, V., Lebeer, S., De Keersmaecker, S., Vanderleyden, J., Gruber, H. J. and Dufrene, Y. F. (2009) Stretching polysaccharides on live cells using single molecule force spectroscopy. *Nat Protoc* **4**(6): 939–946.
· protocol for force studies of cell-surface polysaccharides by atomic force microscopy and description by the freely-jointed chain model

Gore, J., Bryant, Z., Stone, M. D., Nollmann, M., Cozzarelli, N. R. and Bustamante, C. (2006) Mechanochemical analysis of DNA gyrase using rotor bead tracking. *Nature* **439**(7072): 100–104.
· investigation of topological changes in DNA and mechanochemical coupling during negative supercoiling by DNA gyrase using the magnetic tweezers rotor-bead assay

Hecht, E., Knittel, P., Felder, E., Dietl, P., Mizaikoff, B. and Kranz, C. (2012) Combining atomic force-fluorescence microscopy with a stretching device for analyzing mechanotransduction processes in living cells. *Analyst* **137**(22): 5208–5214.
· combination of AFM with fluorescence detection on isolated molecules and in living cells

Heller, I., Hoekstra, T. P., King, G. A., Peterman, E. J. and Wuite, G. J. (2014) Optical tweezers analysis of DNA-protein complexes. *Chem Rev* **114**(6): 3087–3119.
· review article with a focus on applications to DNA-protein interactions

Herbert, K. M., La Porta, A., Wong, B. J., Mooney, R. A., Neuman, K. C., Landick, R. and Block, S. M. (2006) Sequence-resolved detection of pausing by single RNA polymerase molecules. *Cell* **125**(6): 1083–1094.
· optical tweezers study on sequence-dependent pausing of RNA polymerase

Hoffmann, T. and Dougan, L. (2012) Single molecule force spectroscopy using polyproteins. *Chem Soc Rev* **41**(14): 4781–4796.
· review article highlighting insights into protein folding by atomic force microscopy

Janovjak, H., Muller, D. J. and Humphris, A. D. (2005) Molecular force modulation spectroscopy revealing the dynamic response of single bacteriorhodopsins. *Biophys J* **88**(2): 1423–1431.
· stepwise unfolding of secondary structure elements of bacteriorhodopsin in native purple membranes

Janshoff, A., Neitzert, M., Oberdorfer, Y. and Fuchs, H. (2000) Force spectroscopy of molecular systems-single molecule spectroscopy of polymers and biomolecules. *Angew Chem Int Ed* **39**(18): 3212–3237.
· review article on mechanochemical models for stretching of biopolymers, providing a theoretical treatment of the worm-like chain, freely-jointed chain and Gaussian chain models

Jarzynski, C. (1997) Nonequilibrium equality for free energy differences. *Phys Rev Lett* **78**(14): 2690–2693.
· derivation of the calculation of ΔG from the distribution of rupture forces and the distribution of work values determined over many experiments

Kalle, W. and Strappe, P. (2012) Atomic force microscopy on chromosomes, chromatin and DNA: A review. *Micron* **43**(12): 1224–1231.
· review article about applications of AFM to image DNA and DNA-protein complexes

Kedrov, A., Janovjak, H., Sapra, K. T. and Muller, D. J. (2007) Deciphering molecular interactions of native membrane proteins by single-molecule force spectroscopy. *Annu Rev Biophys Biomol Struct* **36**: 233–260.
· review article about force-induced unfolding of membrane proteins by atomic force microscopy

Kedrov, A., Ziegler, C., Janovjak, H., Kuhlbrandt, W. and Muller, D. J. (2004) Controlled unfolding and refolding of a single sodium-proton antiporter using atomic force microscopy. *J Mol Biol* **340**(5): 1143–1152.
· mechanical unfolding study of the bacterial Na+/H+ antiporter NhaA

Kodera, N., Yamamoto, D., Ishikawa, R. and Ando, T. (2010) Video imaging of walking myosin V by high-speed atomic force microscopy. *Nature* **468**(7320): 72–76.
· imaging of myosin V movement along actin by high-speed AFM

Koster, D. A., Palle, K., Bot, E. S., Bjornsti, M. A. and Dekker, N. H. (2007) Antitumour drugs impede DNA uncoiling by topoisomerase I. *Nature* **448**(7150): 213–217.
· probing the mechanism of topoisomerase I poisoning by the drug camptothecin using magnetic tweezers

Lang, M. J., Fordyce, P. M., Engh, A. M., Neuman, K. C. and Block, S. M. (2004) Simultaneous, coincident optical trapping and single-molecule fluorescence. *Nat Methods* **1**(2): 133–139.
· combination of optical tweezers with fluorescence detection and TIRF microscopy

Lee, S. and Hohng, S. (2013) An optical trap combined with three-color FRET. *J Am Chem Soc* **135**(49): 18260–18263.
· combination of optical tweezers with three-color fluorescence detection for three-color FRET

Li, P. T., Collin, D., Smith, S. B., Bustamante, C. and Tinoco, I., Jr. (2006) Probing the mechanical folding kinetics of TAR RNA by hopping, force-jump, and force-ramp methods. *Biophys J* **90**(1): 250–260.
· comparative study of force-induced unfolding of HIV TAR RNA using force hopping, force jump- and force-ramp optical tweezers experiments

Liphardt, J., Dumont, S., Smith, S. B., Tinoco, I., Jr. and Bustamante, C. (2002) Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science* **296**(5574): 1832–1835.
· experimental proof of the Jarzynski equation that allows calculation of ΔG values from non-equilibrium unfolding

Liphardt, J., Onoa, B., Smith, S. B., Tinoco, I., Jr. and Bustamante, C. (2001) Reversible unfolding of single RNA molecules by mechanical force. *Science* **292**(5517): 733–737.

· force-induced unfolding studies on the folding pathway of the P5abc region of the *Tetrahymena* ribozyme

Long, X., Parks, J. W., Bagshaw, C. R. and Stone, M. D. (2013) Mechanical unfolding of human telomere G-quadruplex DNA probed by integrated fluorescence and magnetic tweezers spectroscopy. *Nucleic Acids Res* **41**(4): 2746–2755.
· analysis of mechanical unfolding of telomeric G-quadruplex sequences by combining magnetic tweezers with FRET

Ma, J., Bai, L. and Wang, M. D. (2013) Transcription under torsion. *Science* **340**(6140): 1580–1583.
· investigation of the effect of DNA supercoiling on transcription by RNA polymerases transcription with optical tweezers and torque

Marko, J. F. and Siggia, E. D. (1995) Stretching DNA. *Macromolecules* **28**(26): 8759–8770.
· statistical mechanics treatment of the worm-like chain model and its applicability to mechanical unfolding of DNA

Michaelis, J. and Treutlein, B. (2013) Single-molecule studies of RNA polymerases. *Chem Rev* **113**(11): 8377–8399.
· review article summarizing the insight into the mechanism of transcription by RNA polymerase obtained from a variety of single-molecule approaches

Moffitt, J. R., Chemla, Y. R., Smith, S. B. and Bustamante, C. (2008) Recent advances in optical tweezers. *Annu Rev Biochem* **77**: 205–228.
· review article about applications and limitations of optical tweezers

Moller, C., Fotiadis, D., Suda, K., Engel, A., Kessler, M. and Muller, D. J. (2003) Determining molecular forces that stabilize human aquaporin-1. *J Struct Biol* **142**(3): 369–378.
· investigation of the folding pathways of the water channel aquaporin by atomic force microscopy

Molloy, J. E., Burns, J. E., Kendrick-Jones, J., Tregear, R. T. and White, D. C. (1995) Movement and force produced by a single myosin head. *Nature* **378**(6553): 209–212.
· optical tweezers study of the force and step size during myosin movement along actin filaments

Muller, S. A., Muller, D. J. and Engel, A. (2011) Assessing the structure and function of single biomolecules with scanning transmission electron and atomic force microscopes. *Micron* **42**(2): 186–195.
· combination of AFM imaging with force mapping and force spectroscopy at individual points

Neuman, K. C., Charvin, G., Bensimon, D. and Croquette, V. (2009) Mechanisms of chiral discrimination by topoisomerase IV. *Proc Natl Acad Sci U S A* **106**(17): 6986–6991.
· magnetic tweezers study on the relaxation of positive and negative DNA supercoils by topoisomerase IV

Neuman, K. C. and Nagy, A. (2008) Single-molecule force spectroscopy: Optical tweezers, magnetic tweezers and atomic force microscopy. *Nat Methods* **5**(6): 491–505.
· review article that compares optical and magnetic tweezers and atomic force microscopy

Nollmann, M., Stone, M. D., Bryant, Z., Gore, J., Crisona, N. J., Hong, S. C., Mitelheiser, S., Maxwell, A., Bustamante, C. and Cozzarelli, N. R. (2007) Multiple modes of *Escherichia coli* DNA gyrase activity revealed by force and torque. *Nat Struct Mol Biol* **14**(4): 264–271.
· study on the response of different DNA gyrase activities to external forces by magnetic tweezers

Ogawa, T., Yogo, K., Furuike, S., Sutoh, K., Kikuchi, A. and Kinosita, K., Jr. (2015) Direct observation of DNA overwinding by reverse gyrase. *Proc Natl Acad Sci U S A* **112**(24): 7495–7500.
· monitoring positive DNA supercoiling by reverse gyrase with magnetic tweezers

Qu, X., Wen, J. D., Lancaster, L., Noller, H. F., Bustamante, C. and Tinoco, I., Jr. (2011) The ribosome uses two active mechanisms to unwind messenger RNA during translation. *Nature* **475**(7354): 118–121.
· optical tweezers study of ribosome-mediated RNA unwinding

Ribeck, N. and Saleh, O. A. (2008) Multiplexed single-molecule measurements with magnetic tweezers. *Rev Sci Instrum* **79**(9): 094301.
· multiplexing of magnetic tweezers experiments

Rief, M., Clausen-Schaumann, H. and Gaub, H. E. (1999) Sequence-dependent mechanics of single DNA molecules. *Nat Struct Biol* **6**(4): 346–349.
· study of double-stranded DNA stability as a function of sequence by atomic force microscopy

Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M. and Gaub, H. E. (1997) Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**(5315): 1109–1112.
· multi-step unfolding of a multi-domain protein by atomic force microscopy

Sapra, K. T., Besir, H., Oesterhelt, D. and Muller, D. J. (2006) Characterizing molecular interactions in different bacteriorhodopsin assemblies by single-molecule force spectroscopy. *J Mol Biol* **355**(4): 640–650.
· atomic force microscopy study of the effect of dimer and trimer formation on the stability of bacteriorhodopsin

Selvin, M. and Ha, T. (2008) *Single-Molecule Techniques – A Laboratory Manual*. New York, Cold Spring Harbour Press.
· an overview about single-molecule techniques including force methods with a focus on practical aspects and protocols.

Simmons, R. (1996) Molecular motors: Single-molecule mechanics. *Curr Biol* **6**(4): 392–394.
· review article on optical tweezer studies on myosin, kinesin, and RNA polymerase

Smith, S. B., Cui, Y. and Bustamante, C. (1996) Overstretching B-DNA: The elastic response of individual double-stranded and single-stranded DNA molecules. *Science* **271**(5250): 795–799.
· mechanochemical study of single- and double-stranded DNA

Strick, T. R., Croquette, V. and Bensimon, D. (2000) Single-molecule analysis of DNA uncoiling by a type II topoisomerase. *Nature* **404**(6780): 901–904.
· magnetic tweezers study of DNA relaxation by DNA topoisomerase II

Svoboda, K. and Block, S. M. (1994) Force and velocity measured for single kinesin molecules. *Cell* **77**(5): 773–784.
· mechanochemical study of kinesin movement along microtubules

Swoboda, M., Grieb, M. S., Hahn, S. and Schlierf, M. (2014) Measuring two at the same time: Combining magnetic tweezers with single-molecule FRET. *EXS* **105**: 253–276.
· review article about combining magnetic tweezers with fluorescence detection and FRET

Tinoco, I., Jr., Li, P. T. and Bustamante, C. (2006) Determination of thermodynamics and kinetics of RNA reactions by force. *Q Rev Biophys* **39**(4): 325–360.
· review article on the analysis of the thermodynamics and kinetics of RNA folding by optical tweezers, including a detailed treatment of calibration procedures

Uchihashi, T., Iino, R., Ando, T. and Noji, H. (2011) High-speed atomic force microscopy reveals rotary catalysis of rotorless F(1)-ATPase. *Science* **333**(6043): 755–758.
· real-time imaging of the rotation of F1 ATPase by high-speed AFM at sub-second resolution

Veigel, C., Coluccio, L. M., Jontes, J. D., Sparrow, J. C., Milligan, R. A. and Molloy, J. E. (1999) The motor protein myosin-I produces its working stroke in two steps. *Nature* **398**(6727): 530–533.
· analysis of the mechanochemical cycle of actomyosin by optical tweezers

Visscher, K., Schnitzer, M. J. and Block, S. M. (1999) Single kinesin molecules studied with a molecular force clamp. *Nature* **400**(6740): 184–189.
· force study of the movement of myosin along actin filaments

Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J. and Block, S. M. (1998) Force and velocity measured for single molecules of RNA polymerase. *Science* **282**(5390): 902–907.
· mechanochemical study of the response of RNA polymerase translocation to forces using optical tweezers

Wuite, G. J., Smith, S. B., Young, M., Keller, D. and Bustamante, C. (2000) Single-molecule studies of the effect of template tension on T7 DNA polymerase activity. *Nature* **404**(6773): 103–106.
· optical tweezers study of the effect of forces on DNA polymerase activity

Yin, H., Wang, M. D., Svoboda, K., Landick, R., Block, S. M. and Gelles, J. (1995) Transcription against an applied force. *Science* **270**(5242): 1653–1657.
· optical tweezers study of transcription by RNA polymerase under force

Zhou, R., Schlierf, M. and Ha, T. (2010) Force-fluorescence spectroscopy at the single-molecule level. *Methods Enzymol* **475**: 405–426.
· combination of optical tweezers with confocal microscopy and fluorescence detection

Zlatanova, J. and Leuba, S. H. (2003) Magnetic tweezers: A sensitive tool to study DNA and chromatin at the single-molecule level. *Biochem Cell Biol* **81**(3): 151–159.
· review article highlighting the applications of magnetic tweezers to DNA and DNA-dependent enzymes and its potential for studies on chromatin

28

Transient kinetic methods

In Part II, we derived rate laws for different types of reactions and formally treated pre-steady state and steady-state kinetics. We have seen before that kinetic information can be derived by observing single molecules fluctuate between different states (Sections 12, 25.7.4). While these experiments do not require synchronization of molecules, in transient kinetic experiments a reaction is typically started by manual or automated mixing of the components. If the disappearance of reactants or appearance of intermediates and products is accompanied by a change in a spectroscopic signal, the reaction is monitored continuously in real time. Under certain conditions, reactions can also be followed by the heat release or uptake using calorimetry (see Section 30.1.6). Reactions that are not associated with a change in spectroscopic signal are analyzed discontinuously, by stopping the reaction at different time points, followed by quantification of reactant(s) and/or product(s) at each time point.

This chapter focuses on the main experimental techniques to initiate reactions of different velocities. Manual mixing typically requires a few seconds before the first data point or sample can be taken, which is the *dead time* of the experiment. Slow reactions will not have progressed much during this dead time, but fast reactions might already have completed before the measurement has even started or the first sample has been taken. In simple words: for rapid reactions, rapid mixing is essential.

28.1 STOPPED FLOW

A *stopped flow* instrument (Figure 28.1) contains the reactants in two syringes that both connect to a mixing or reaction chamber. The reaction is started by movement of a plate, either driven by expansion of compressed gas or by a motor, which rapidly pushes the pistons of the syringes. The reactants flow into the mixing chamber where they displace the old solution from the previous run into a stop syringe. As soon as the piston of the stop syringe contacts a stop plate, the mixing process is finished and the measurement starts. The overall dead times achieved by stopped flow instruments are about 1 ms. The reaction progress is typically followed *via* absorption, fluorescence intensity or anisotropy, or circular dichroism. Any spectroscopic signal that is a measure of concentration can be used to monitor the reaction, even NMR. From the time dependence of the spectroscopic signal, the rate constant for the reaction may be determined (see Chapters 6–8). By varying the concentrations of reactants, the underlying kinetic mechanism can be deduced. Rate constants can be determined as a function of temperature to determine activation energies (Chapter 15).

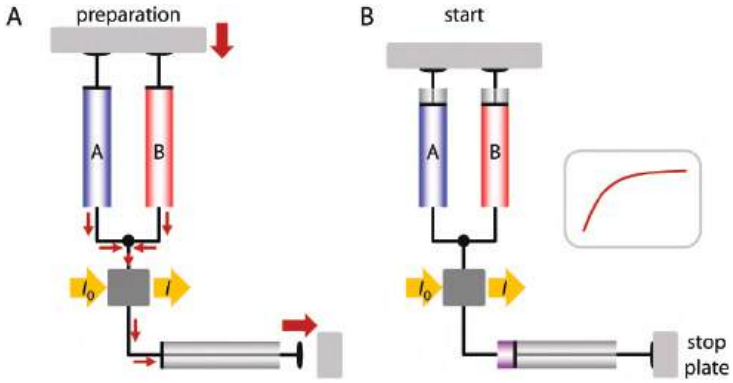


FIGURE 28.1: Single-mixing stopped flow instrument. A: The top plate pushes the syringes containing the reactants A (blue) and B (red), leading to rapid mixing in the mixing chamber (black circle). B: The mixing process ends when the solution pushed into the mixing and observation chamber (gray) reaches the stop syringe (purple), and the plunger hits the stop plate (right). At this point, the measurement of a spectroscopic signal (here: absorption) for the disappearance of reactants or appearance of products is started.

Single-mixing stopped flow instruments have two syringes that are moved simultaneously to mix two reactants. *Double-mixing* stopped flow instruments (Figure 28.2) can realize more complex mixing schemes, such as pre-formation of a complex, followed by mixing of this complex with a third compound. This way, transient intermediates can be populated and their reactions with other compounds can be studied.

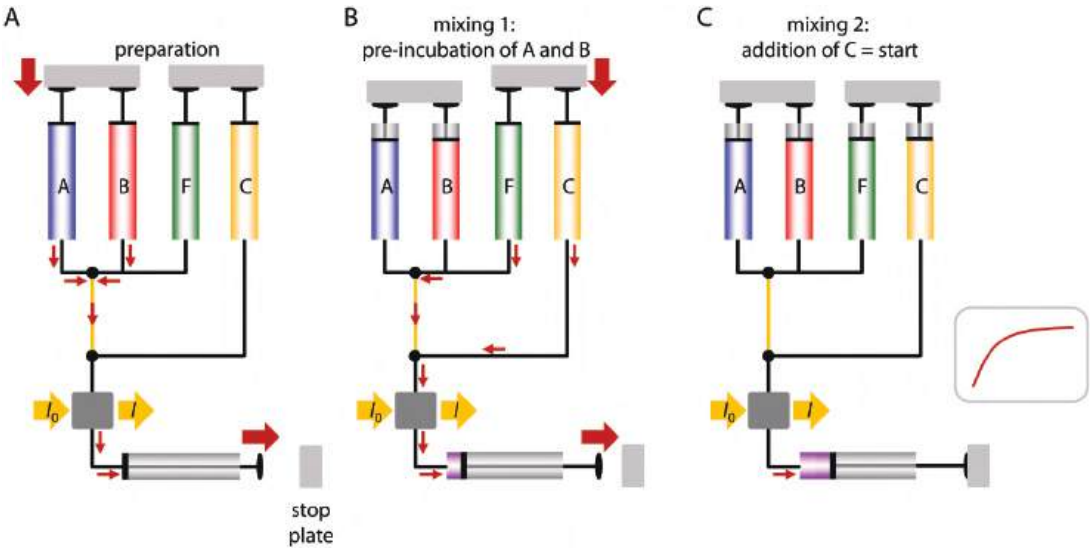


FIGURE 28.2: Double-mixing stopped flow instrument. A: A double-mixing stopped flow typically consists of four syringes, and has two plates to move the syringes independently in pairs. B: In the first mixing event, plate 1 pushes the two syringes containing reactants A and B, leading to mixing of A and B, and allowing for a reaction, such as AB complex formation, in the pre-incubation loop (orange). C: After the pre-incubation time, plate 2 pushes syringes F (for flush) and C, leading to mixing of C and AB. Syringe F is filled with buffer and pushes the aged mixture of A and B forward. The measurement starts when the plunger of the stop syringe hits the stop plate at the end of the second mixing event.

Double-mixing experiments are often used in protein folding studies, where the protein of interest is first unfolded for a certain time, and then brought back into native conditions to initiate refolding. Alternatively, the unfolded protein is refolded for a certain time, and then again transferred to denaturing conditions. From the number of phases in the refolding/unfolding reactions and their amplitudes, information on folding intermediates and folding pathways can be deduced (Box 28.1).

BOX 28.1: DOUBLE-MIXING STOPPED FLOW EXPERIMENTS TO ANALYZE PATHWAYS OF PROTEIN FOLDING.

Ribonuclease T_1 has been used as a model system to study the role of proline isomerization in protein folding. In the native state of RNase T_1 the peptide bonds preceding Pro39 and Pro55 are in the *cis*-configuration. Equilibrium unfolding experiments of RNase T_1 are consistent with the two-state model (Section 18.5.4), but the folding and unfolding reactions are more complex. After loss of the native structure, the *cis*-prolyl peptide bonds isomerize to the more stable *trans* configuration. As a result, the unfolded state of RNase T_1 is an equilibrium mixture of four species with different configurations of the prolyl peptide bonds: 39*cis*/55*cis*, 39*cis*/55*trans*, 39*trans*/55*cis*, and 39*trans*/55*trans* (Figure 28.3). 39*cis*/55*cis* folds directly into the native state, whereas all other species rapidly form partially folded intermediates that reach the native state through slow *trans-cis* isomerization of one or two prolines. Double-mixing stopped flow experiments were used to first populate the unfolded state 39*cis*/55*cis* that is only present to 2–4% in equilibrium by short unfolding for a time Δt_u, and to then monitor its refolding by an increase in fluorescence. Refolding curves show double-exponential behavior, with a fast phase (k = 7.6 s^{-1}) and a slow phase (k = 1 s^{-1}). The amplitudes of the two phases depend on the unfolding time Δt_u in the first step that determines the populations of the different prolyl isomers. Both amplitudes increase with increasing unfolding time, reach a maximum, and then decrease to a small final value (Figure 28.3). This behavior is characteristic of consecutive reactions (Section 8.3), and the curves provide the rate constants of production and decay for the species that refolds in each phase (Figure 28.3). The rapid refolding phase was assigned to the refolding of the unfolded RNase T_1 with 39*cis*/55*cis* configuration, generated by unfolding of N and decaying to other species by prolyl isomerization. The second phase was assigned to one or several species that are formed from 39*cis*/55*cis* by prolyl isomerization and converted into other species also by prolyl isomerization. In conjunction with folding data of variants with substitutions at Pro39 and Pro55, the rate constants for prolyl isomerization of the peptide bonds preceding Pro39 and Pro55 in the unfolded state were extracted by global analysis of the refolding reactions. Mayr *et al.* (1996) Biochemistry 35(17): 5550–5561.

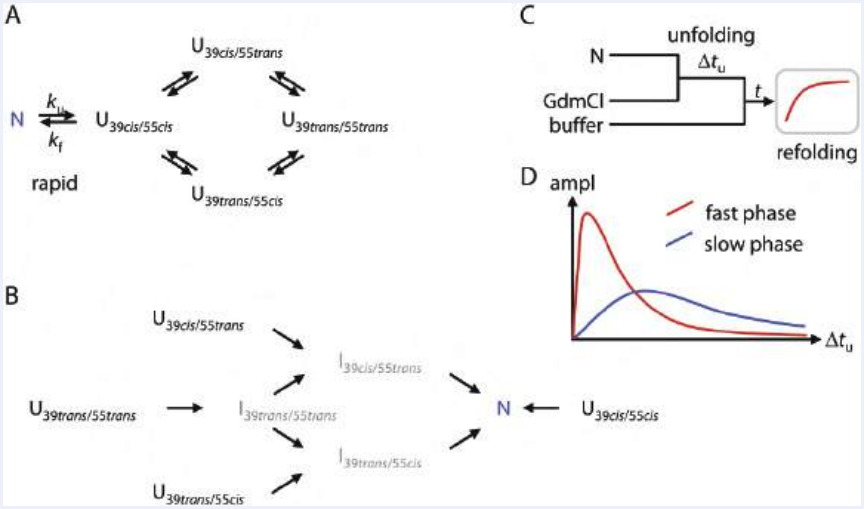


FIGURE 28.3: Analysis of RNase T_1 folding by double-mixing stopped flow experiments. A: Equilibrium unfolding scheme. B: Folding kinetics. U: unfolded, I: intermediate, N: native state. C: Schematic depiction of double-mixing experiments. Native RNase T_1 is unfolded for the time Δt_u by dilution into buffer with 6 M GdmCl, pH 1.6, and refolded by dilution to 1 M GdmCl, pH 4.6. Folding is monitored by fluorescence. D: Amplitude of fast and slow phases of refolding as a function of the unfolding time Δt_u.

28.2 QUENCH FLOW

A quench flow instrument (Figure 28.4) also contains the reactants in two syringes. In addition to the stopped flow instrument, it contains a third syringe with a quenching solution that stops the reaction of interest at a certain time point. Quench flow is thus a discontinuous method to monitor the progress of a reaction. To initiate the reaction, the motor plate pushes all three syringes, such that the reactants are mixed and passed through a reaction loop. The quenching solution from the third syringe is made to travel a longer path. The quench solution is combined with the reaction mixture only after the reaction mixture has passed the reaction loop, and the reaction time is thus determined by the volume of the reaction loop. Using reaction loops of different volumes, the reaction can be stopped at different time points. The individual samples are expelled from the instrument into a collection vial and subjected to analysis by any suitable method, such as thin layer chromatography, HPLC, quantitative mass spectrometry, *etc.* The dead time, *i.e.* the shortest reaction time that can be realized in a quench flow instrument, is about 1–2 ms. Longer reaction times above the travel time through the ageing loop can be realized by a push-wait-push sequence: a first push of the mixing plate mixes A and B, the reaction can proceed during the waiting time, and the second push adds the quench solution and expels the sample from the instrument. On the one hand, quench flow is more versatile than stopped flow because no spectroscopic signal is needed. On the other hand, each sample has to be analyzed separately, which is cumbersome and yields fewer data points. Again, more complex implementations of double mixing followed by quenching allow the realization of sequential reactions.

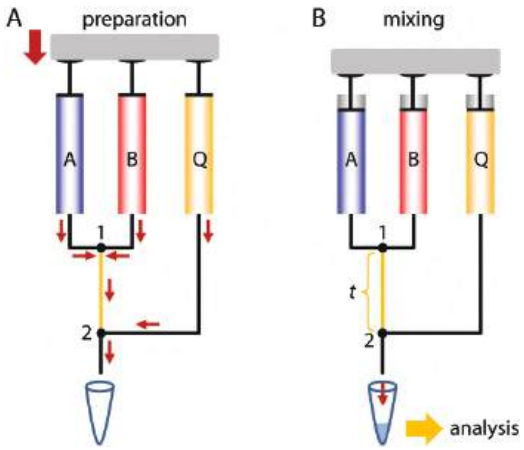


FIGURE 28.4: Rapid mixing by quench flow methods. A: A quench flow system consists of three syringes, filled with the reactants A and B and the quencher compound Q that stops the reaction. A moving plate pushes all three syringes simultaneously, leading to mixing of A and B in the first mixing chamber (black circle, 1). The quencher is added later in the path in mixer 2 (black circle, 2) by a second push after a delay time. B: During the time the solution passes the reaction loop (orange), the reaction between A and B occurs. When the reaction mixture reaches mixer 2, addition of the quench solution stops the reaction. The length of the reaction loop thus determines the reaction time, and quench flow instruments contain several reaction loops of different lengths to stop the reaction at different time points. The sample is expelled from the instrument, collected, and subjected to analysis for reactant and/or product content.

Quench flow experiments are used for reactions that cannot be followed by a spectroscopic signal in real time. Quench flow analysis has been used to study the kinetic mechanism of primer/template binding and the incorporation of nucleotides by HIV reverse transcriptase, and the stimulation of ATP hydrolysis by the chaperone DnaK in the presence of its co-chaperone DnaJ (Box 28.2).

BOX 28.2: QUENCH FLOW EXPERIMENTS TO DETERMINE THE STIMULATION OF ATP HYDROLYSIS BY CHAPERONES IN THE PRESENCE OF CO-CHAPERONES.

The chaperone DnaK assists protein folding by interacting with exposed hydrophobic regions of unfolded proteins, preventing their aggregation. DnaK is an ATP-binding protein. In the ATP state, binding and dissociation of peptides is rapid, but the peptide affinity is low. ATP hydrolysis converts DnaK into the ADP state that shows slow peptide binding and dissociation, but has a high peptide affinity. Dissociation of ADP and re-binding of ATP starts a new catalytic cycle. The cycle of ATP hydrolysis is the timer for binding and release of peptides, and central for the chaperone function of DnaK. The rate of ATP hydrolysis is increased in the presence of the co-chaperone DnaJ, the rate of nucleotide exchange is stimulated by the co-chaperone GrpE. The stimulation of ATP hydrolysis by DnaJ was determined in quench flow experiments. The DnaK-ATP complex was pre-formed during 10 s, and subsequently mixed with DnaJ in a quench flow apparatus (Figure 28.5). After different reaction times t, the reaction was stopped by addition of perchloric acid, ATP and phosphate were separated by thin layer chromatography, and the amount of ATP was quantified. The rate constant of ATP hydrolysis in the presence of DnaJ was $k = 0.14\ s^{-1}$, which is 100-fold higher than the rate constant for ATP hydrolysis by DnaK alone ($k = 1.5{\cdot}10^{-3}\ s^{-1}$). The maximal stimulation was determined by performing the same reaction at different concentrations of DnaJ. The maximal rate constant was $k = 0.79\ s^{-1}$, corresponding to a stimulation by 530-fold. From the plot of the rate constant as a function of the DnaJ concentration, a K_d value of 16 μM for dissociation of DnaJ from the DnaK-ATP complex was obtained. Laufen *et al.* (1999) Proc. Natl. Acad. Sci. U S A 96(10): 5452–5457.

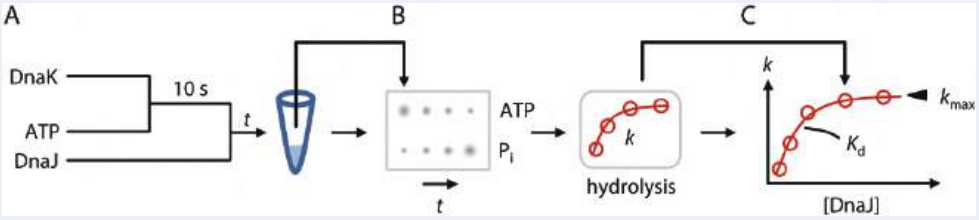


FIGURE 28.5: Stimulation of ATP hydrolysis by DnaK in the presence of DnaJ analyzed by quench flow. A: Schematic depiction of a double-mixing experiment to preincubate DnaK with ATP, with subsequent mixing of the DnaK-ATP complex with DnaJ. B: The reaction is stopped at times t, and the ATP and inorganic phosphate (P_i) content is analyzed by thin layer chromatography to obtain the rate constant of hydrolysis. C: The rate constant as a function of DnaJ concentration provides the rate constant at maximal stimulation, k_{max}, and the K_d value for dissociation of DnaJ from the DnaK-ATP complex.

28.3 LASER FLASH PHOTOLYSIS

For very rapid reactions that occur on the millisecond time scale or faster, starting the reaction by mechanical mixing is not fast enough. Light-sensitive processes, such as the *cis/trans* isomerization of retinal in rhodopsins or the reactions of other photoreceptors, can be triggered on the microsecond time scale by illumination. Certain processes can be initiated rapidly by illumination although they are not naturally triggered by light, such as the dissociation of CO from heme (see Section 24.5). Many processes can be rendered light-sensitive using *caged compounds*. These reagents are inactivated by photolabile protective groups that can be removed by illumination with (typically) UV light. Removal of the protective group liberates the active component and starts the reaction

in situ. The reaction can then be followed continuously using spectroscopic probes. The dead time of these procedures is on the order of 1 µs. Examples of caged compounds include caged ATP (Figure 28.6) for the studies of ATPases, such as the muscle protein myosin, and caged GTP to investigate GTPases, such as G-proteins.

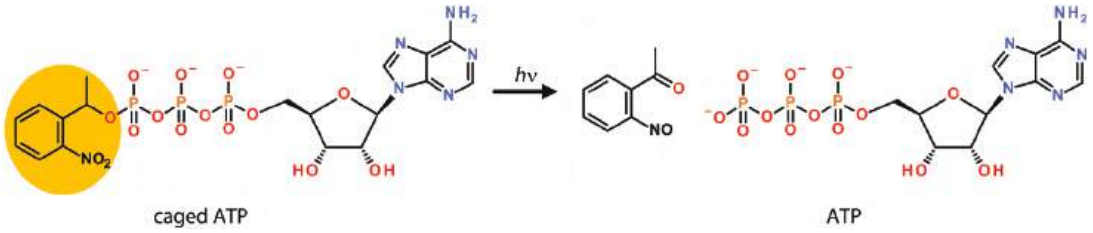


FIGURE 28.6: Activation of caged compounds. In caged ATP (1-(2-nitrophenylethyl)-ATP, NPE-ATP), the γ-phosphate is protected by a 2-nitrophenylethyl group (orange). In this state, ATP is not hydrolyzed by ATPases. The NPE group has an extinction coefficient of 660 M^{-1} cm^{-1} at 347 nm, and the caged compound is activated by photo-induced cleavage of the group ($h\nu$) to rapidly liberate ATP.

Caged Ca^{2+} allows the rapid release of Ca^{2+} to trigger signaling cascades. Nitrophenyl-EGTA (Figure 28.7) cages Ca^{2+} by high-affinity binding. Photolysis cleaves the chelator in two, leading to a drop in Ca^{2+} affinity and its release. Caged glutamate (4-methoxy-7-nitro-indolinyl-Glu; Figure 28.7) does not bind to postsynaptic glutamate-gated ion channels. Illumination at 350 nm (ε = 4300 M^{-1} cm^{-1}) liberates glutamate that acts as a neurotransmitter and activates ion channels.

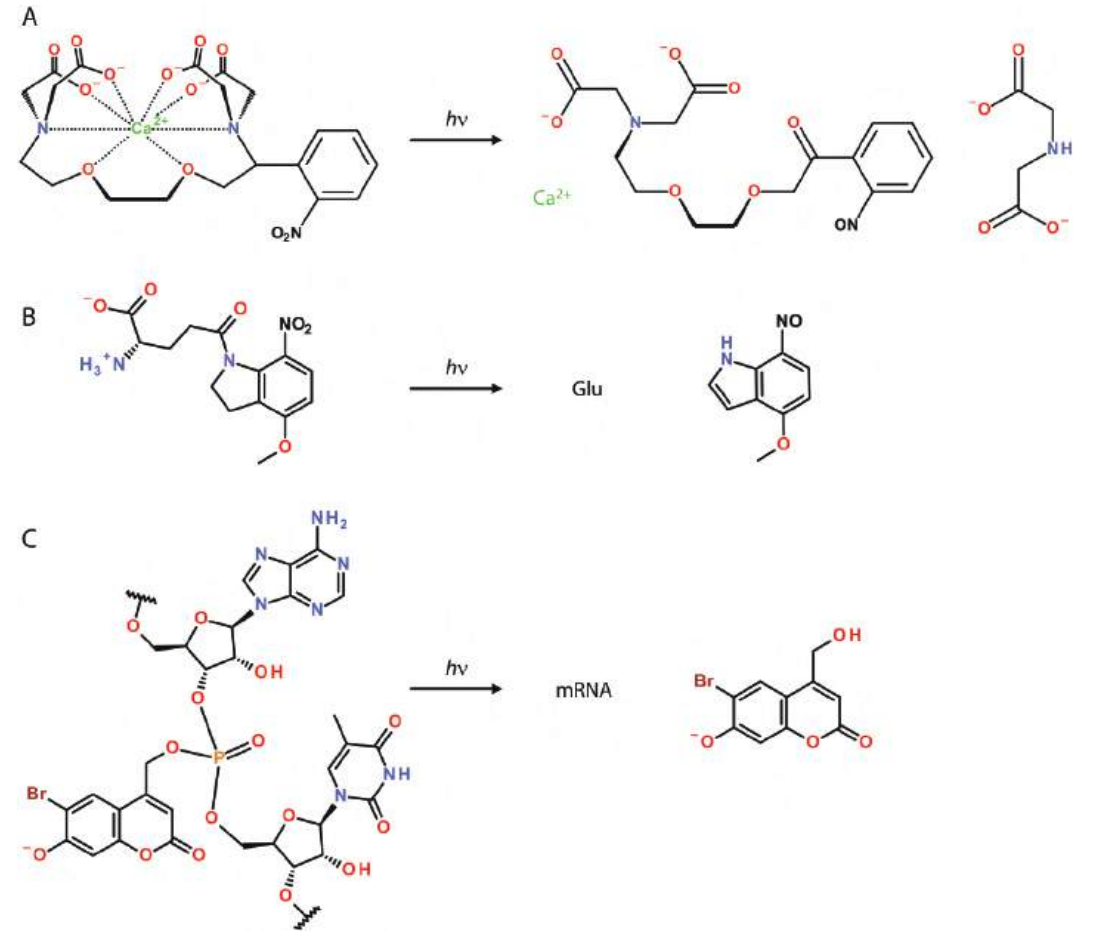


FIGURE 28.7: Examples of caged compounds. A: Caged Ca^{2+}, a complex of nitrophenyl-EGTA and Ca^{2+}. B: Caged glutamate (4-methoxy-7-nitro-indolinyl-glutamate). C: Caged mRNA, protected by the bromohydroxycoumarin group at the phosphate groups.

Larger molecules can also be inactivated by caging. In caged mRNA (Figure 28.7), the backbone phosphates are protected by the bromohydroxycoumarin group that is introduced during chemical synthesis. Caged mRNA is activated by photocleavage of the bromohydroxycoumarin group and then serves as a template for protein biosynthesis. It is also possible to cage individual groups in proteins. Using chemically synthesized aminoacyl-tRNAs that carry photocleavable o-nitrophenyl-modified phosphoserine, phosphothreonine, or phosphotyrosine, caged phosphoproteins can be generated by *in vitro* translation and suppression of stop codons. Using suppression of stop codons, caged cysteine and caged tyrosine (modified with the photocleavable o-nitrobenzyl group) have been incorporated into the nicotinic acetylcholine receptor *in vivo*. By photocleavage of the o-nitrobenzyl groups, active receptor molecules were recovered with a time constant of 500 ms, which is slower than the photoreaction (1 ms) and most likely caused by a conformational change of the receptor.

28.4 RELAXATION KINETICS: PRESSURE- AND TEMPERATURE-JUMP

Relaxation kinetics is based on a small, but rapid, disturbance of a reaction in equilibrium, either by applying a jump in temperature or in pressure. Due to the temperature and pressure dependence of equilibrium constants (see Section 2.1), the position of the equilibrium will be different after the disturbance. The system therefore relaxes to the new equilibrium in a process that depends on the rate constants for the forward and backward reaction. The relaxation to the new equilibrium is followed continuously using a spectroscopic signal. For conversions of two species, such as folding and unfolding

$$A \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} B \qquad \text{scheme 8.1}$$

the observed relaxation time τ, *i.e.* the time constant for reaching the new equilibrium position, is

$$\tau = \frac{1}{k_1 + k_{-1}} \qquad \text{eq. 28.1}$$

The constants k_1 and k_{-1} are the rate constants at the temperature or pressure after the jump. For a simple bimolecular reaction of

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AB \qquad \text{scheme 8.2}$$

the observed relaxation time τ for the approach of the new equilibrium is

$$\tau = \frac{1}{k_1\left([A]+[B]\right) + k_{-1}} \qquad \text{eq. 28.2}$$

Thus, from the relaxation times measured at different concentrations of A and B, the forward (association) rate constant k_1 and the reverse (dissociation) rate constant k_{-1} can be determined. *Temperature jumps* are usually achieved by illumination with infrared laser light. A temperature jump by 5–10°C can easily be completed within 0.1–1 µs. *Pressure jumps* in a pressure chamber can be achieved by a quick release valve or by bursting of a membrane. While these approaches are limited to pressure drops, modern instruments achieve pressure changes in both directions by using piezoelectric actuators. This approach offers the additional advantage of performing a large number of repetitions of the experiment. Accumulation of the measured transients significantly improves the signal-to-noise ratio. The dead time of such a pressure jump instrument is on the order of 50–100 µs.

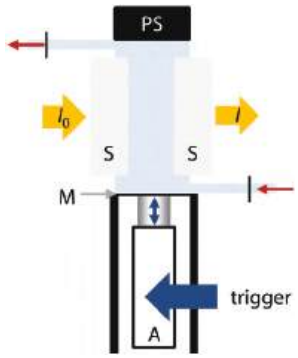


FIGURE 28.8: Principle of a pressure-jump instrument. An electric trigger pulse to the piezo actuator (A) causes its expansion and moves the piston against the membrane (M) that delimits the sample chamber to the bottom (voltage increase/pressure increase) or away from it (voltage decrease/pressure decrease; double-headed blue arrow). A pressure sensor (PS) is attached to the top of the sample chamber. Valves at the inlet and outlet (red arrows) enable changing of the sample. The change in absorption (I_0, I, orange) during relaxation to the new equilibrium after the pressure jump can be monitored through sapphire windows (S). Fluorescence can also be used as a probe in pressure jump experiments.

T-jump and pressure-jump relaxation kinetics have been applied to protein folding down to the microsecond time scale (Box 28.3). Relaxation methods can be combined with NMR, SAXS, FT-IR, and CD/ORD.

BOX 28.3: PRESSURE JUMP AND PROTEIN FOLDING.

A number of small proteins fold into their native states on the time scale of milliseconds and faster. The folding of these proteins is complete within the dead time of mixing of stopped flow devices, and therefore cannot be studied with this technique. One example of a rapidly folding protein is the cold-shock protein CspB. Folding and unfolding cycles of CspB were investigated during cycles of pressure release and increase in a pressure-jump instrument. Folding/unfolding was monitored on the time scale from 50 μs to 70 s by the associated increase/decrease in fluorescence, and rate constants for folding and unfolding were determined over a wide range of temperatures. As the folding/unfolding traces provide the sum of the folding and unfolding rates, the equilibrium constant of folding was determined independently in thermal unfolding experiments. From the temperature dependence of rate constants and equilibrium constants, the enthalpies, entropies and heat capacities of activation, and the corresponding thermodynamic parameters for folding and unfolding were determined. From the pressure dependence of the equilibrium constant of folding, the activation volume for folding and unfolding can be calculated. The activation volume was positive for both reactions: the transition state of folding/unfolding has a higher volume than the unfolded or native states. By comparing the thermodynamic parameters of folding and unfolding of wild-type CspB with variants in which exposed phenylalanine residues were replaced by alanines, the effect of these residues on folding and stability was interrogated. Jacob *et al.* (1999) Biochemistry 38(10): 2882–2891.

QUESTIONS

28.1 You want to measure GTP hydrolysis by a small G-protein in complex with a regulatory protein. The regulatory protein stimulates the GTPase activity, and the reaction becomes too fast for hand-mixing. How can you do that (1) using stopped flow or (2) using quench flow?

28.2 Why is pressure jump useful to study protein folding?

REFERENCES

Ando, H., Furuta, T., Tsien, R. Y. and Okamoto, H. (2001) Photo-mediated gene activation using caged RNA/DNA in zebrafish embryos. *Nat Genet* **28**(4): 317–325.
· triggering gene expression by flash photolysis of caged mRNA in zebrafish embryos

Barends, T. R., Foucar, L., Ardevol, A., Nass, K., Aquila, A., …, Schlichting, I. (2015) Direct observation of ultrafast collective motions in CO myoglobin upon ligand dissociation. *Science* **350**(6259): 445–450.
· flash photolysis of the Fe-CO bond in myoglobin and structural studies on ultrafast protein conformational changes

Corrie, J. E., DeSantis, A., Katayama, Y., Khodakhah, K., Messenger, J. B., Ogden, D. C. and Trentham, D. R. (1993) Postsynaptic activation at the squid giant synapse by photolytic release of L-glutamate from a 'caged' L-glutamate. *J Physiol* **465**: 1–8.
· postsynaptic activation of squid giant synapse using flash photolysis of caged glutamate.

Dumont, C., Emilsson, T. and Gruebele, M. (2009) Reaching the protein folding speed limit with large, sub-microsecond pressure jumps. *Nat Methods* **6**(7): 515–519.
· fluorescence-detected refolding study of λ repressor applying microsecond pressure drops reveals folding without activation barrier

Ellis-Davies, G. C. (2007) Caged compounds: Photorelease technology for control of cellular chemistry and physiology. *Nat Methods* **4**(8): 619–628.
· review about photocaging and triggering of biochemical reactions by flash photolysis

Gruenewald, B. and Knoche, W. (1978) Pressure jump method with detection of optical rotation and circular dichroism. *Rev Sci Instrum* **49**(6): 797.
· pressure-jump instrument with circular dichroism/optical rotation dispersion detection

Herberhold, H. and Winter, R. (2002) Temperature- and pressure-induced unfolding and refolding of ubiquitin: A static and kinetic Fourier transform infrared spectroscopy study. *Biochemistry* **41**(7): 2396–2401.
· temperature-dependent folding study of ubiquitin using pressure-jump and Fourier transform infrared spectroscopy on the time scale of seconds to minutes

Jacob, M., Holtermann, G., Perl, D., Reinstein, J., Schindler, T., Geeves, M. A. and Schmid, F. X. (1999) Microsecond folding of the cold shock protein measured by a pressure-jump technique. *Biochemistry* **38**(10): 2882–2891.
· pressure-jump analysis of folding of the protein CspB on the microsecond time scale

Kremer, W., Arnold, M., Munte, C. E., Hartl, R., Erlach, M. B., Koehler, J., Meier, A. and Kalbitzer, H. R. (2011) Pulsed pressure perturbations, an extra dimension in NMR spectroscopy of proteins. *J Am Chem Soc* **133**(34): 13646–13651.
· combination of millisecond pressure-jump with NMR spectroscopy

Laufen, T., Mayer, M. P., Beisel, C., Klostermeier, D., Mogk, A., Reinstein, J. and Bukau, B. (1999) Mechanism of regulation of Hsp70 chaperones by DnaJ cochaperones. *Proc Natl Acad Sci U S A* **96**(10): 5452–5457.
· quench flow experiments to determine the stimulation of ATP hydrolysis by the chaperone DnaK in the presence of its co-chaperone DnaJ

Marx, A., Jagla, A. and Mandelkow, E. (1990) Microtubule assembly and oscillations induced by flash photolysis of caged-GTP. *Eur Biophys J* **19**(1): 1–9.
· study of tubulin assembly using caged GTP, flash photolysis, and time-resolved X-ray scattering

Matsumoto, T., Nakagawa, T. and Kuwata, K. (2009) Cold destabilization and temperature jump of the murine prion protein mPrP(23–231). *Biochim Biophys Acta* **1794**(4): 669–673.
· temperature jump study on stability and cold denaturation of prions

Mayr, L. M., Odefey, C., Schutkowski, M. and Schmid, F. X. (1996) Kinetic analysis of the unfolding and refolding of ribonuclease T1 by a stopped-flow double-mixing technique. *Biochemistry* **35**(17): 5550–5561.
· double-jump stopped flow study on prolyl isomerization during folding and unfolding of RNase T1

Panick, G. and Winter, R. (2000) Pressure-induced unfolding/refolding of ribonuclease A: Static and kinetic Fourier transform infrared spectroscopy study. *Biochemistry* **39**(7): 1862–1869.
· pressure-induced unfolding of RNase A studied by Fourier transform infrared spectroscopy on the time scale of minutes

Pearson, D. S., Holtermann, G., Ellison, P., Cremo, C. and Geeves, M. A. (2002) A novel pressure-jump apparatus for the microvolume analysis of protein-ligand and protein-protein interactions: Its application to nucleotide binding to skeletal-muscle and smooth-muscle myosin subfragment-1. *Biochem J* **366**(Pt 2): 643–651.
· report of a pressure jump instrument that achieves pressure increase and decrease by using piezoelectric actuators

Philipson, K. D., Gallivan, J. P., Brandt, G. S., Dougherty, D. A. and Lester, H. A. (2001) Incorporation of caged cysteine and caged tyrosine into a transmembrane segment of the nicotinic ACh receptor. *Am J Physiol Cell Physiol* **281**(1): C195–206.
· incorporation of caged cysteine and caged tyrosine into the nicotinic acetylcholine receptor *in vivo* by stop codon suppression and recovery of active receptors by flash photolysis

Rapp, G., Poole, K. J., Maeda, Y., Guth, K., Hendrix, J. and Goody, R. S. (1986) Time-resolved structural studies on insect flight muscle after photolysis of caged-ATP. *Biophys J* **50**(5): 993–997.
· time-resolved structural study of insect flight muscle after photolysis of caged ATP

Rothman, D. M., Petersson, E. J., Vazquez, M. E., Brandt, G. S., Dougherty, D. A. and Imperiali, B. (2005) Caged phosphoproteins. *J Am Chem Soc* **127**(3): 846–847.
· chemical synthesis of photocaged phosphorylated amino acids and incorporation into nicotinic acetyl choline receptor and the vasodilator-stimulated phosphoprotein VASP by amber stop codon suppression and *in vitro* translation

Schlichting, I. and Goody, R. S. (1997) Triggering methods in crystallographic enzyme kinetics. *Meth Enzymol* **277**: 467–490.
· review on methods to trigger enzymatic reactions in crystals

Schlichting, I., Rapp, G., John, J., Wittinghofer, A., Pai, E. F. and Goody, R. S. (1989) Biochemical and crystallographic characterization of a complex of c-Ha-ras p21 and caged GTP with flash photolysis. *Proc Natl Acad Sci U S A* **86**(20): 7687–7690.
· time-resolved crystallography study of GTP hydrolysis by H-Ras in crystals using caged GTP and flash photolysis

Wirth, A. J., Liu, Y., Prigozhin, M. B., Schulten, K. and Gruebele, M. (2015) Comparing fast pressure jump and temperature jump protein folding experiments and simulations. *J Am Chem Soc* **137**(22): 7152–7159.
· a comparative study of rapid folding of β-sheets using pressure jump, temperature jump, and computational analyses

Woenckhaus, J., Kohling, R., Thiyagarajan, P., Littrell, K. C., Seifert, S., Royer, C. A. and Winter, R. (2001) Pressure-jump small-angle x-ray scattering detected kinetics of staphylococcal nuclease folding. *Biophys J* **80**(3): 1518–1523.
· combined pressure-jump and small-angle X-ray scattering study on folding of staphylococcal nuclease on the seconds to minutes time scale

Wohrl, B. M., Krebs, R., Goody, R. S. and Restle, T. (1999) Refined model for primer/template binding by HIV-1 reverse transcriptase: Pre-steady-state kinetic analyses of primer/template binding and nucleotide incorporation events distinguish between different binding modes depending on the nature of the nucleic acid substrate. *J Mol Biol* **292**(2): 333–344.
· analysis of the kinetic mechanism of primer/template binding and nucleotide incorporation by HIV reverse transcriptase

29

Molecular mass, size, and shape

Accurate mass determination is the key to identify macromolecules and their complexes. Mass spectrometry (MS, Section 29.1) is the only method that determines masses directly. MS relies on the ionization of molecules and the separation of these ions according to their mass in an electric field. Analytical ultracentrifugation (AUC, Section 29.2) measures the sedimentation behavior of macromolecules in a gravitational field, which not only provides information on their mass, but also on their size and shape. In surface plasmon resonance (SPR, Section 29.3), the change in refractive index reports on mass changes due to interactions of binding partners with surface-immobilized molecules. All of these methods are suitable to characterize macromolecular interactions.

29.1 MASS SPECTROMETRY

Mass spectrometry (MS) starts with the ionization of molecular entities in vacuum, followed by the separation of these ions according to molecular mass, and then their detection. A mass spectrum measures the abundance of species with a certain mass to charge ratio, m/z. The determination of molecular masses is the most important application of MS (Section 29.1.5.1). MS can identify non-covalently bound adducts and covalent modifications, and can be used to identify and sequence proteins and nucleic acids (Section 29.1.5.3). Native MS (Section 29.1.5.4) uses gentle ionization protocols and sophisticated ion manipulation methods to determine the composition, stoichiometry, and conformational heterogeneity of macromolecular complexes, giving insight into the biological pathway of complex assembly. Kinetic information down to the millisecond time scale can be gained by measuring the time-dependent exchange of deuterium for hydrogen atoms (Section 29.1.5.2) or measuring exchange of whole protein subunits in multi-protein complexes (Section 29.1.5.4). Finally, MS is often combined with chromatographic techniques, such as reversed phase (RP)-HPLC or size exclusion chromatography, to analyze the components of a mixture one after the other.

29.1.1 Ionization

The first step in MS is ionization of the molecule(s) of interest. Ions are manipulated and analyzed by applying electric fields. There are two main methods for non-destructive ionization of a molecule, *electrospray ionization* (ESI) and *matrix-assisted laser desorption ionization* (MALDI). In MALDI the ionization is done out of the solid state under denaturing conditions whereas ESI can be performed under both denaturing and physiological conditions.

29.1.1.1 Matrix-assisted laser desorption ionization

MALDI vaporizes macromolecules that are embedded in a *matrix*, which is often an organic acid of low molecular mass (Figure 29.1). Sinapinic acid, or *trans*-3,5-dimethoxy-4-hydroxycinnamic acid, is efficient for ionization of proteins, peptides, and lipids, whereas 2,5-dihydroxybenzoic acid and its acetophenone are good options for (glyco)peptides, oligosaccharides, nucleotides, and oligonucleotides. Nucleic acids are also efficiently ionized from 3-hydroxy picolinic acid. For sample preparation, the macromolecule (analyte) in a low ionic strength buffer is mixed with a 10^3-fold excess of the matrix dissolved in a volatile solvent such as acetonitrile/water. Addition of 0.1% trifluoroacetic acid to the solvents promotes protonation of the macromolecule. The mixture is dried and placed under vacuum in the mass spectrometer.

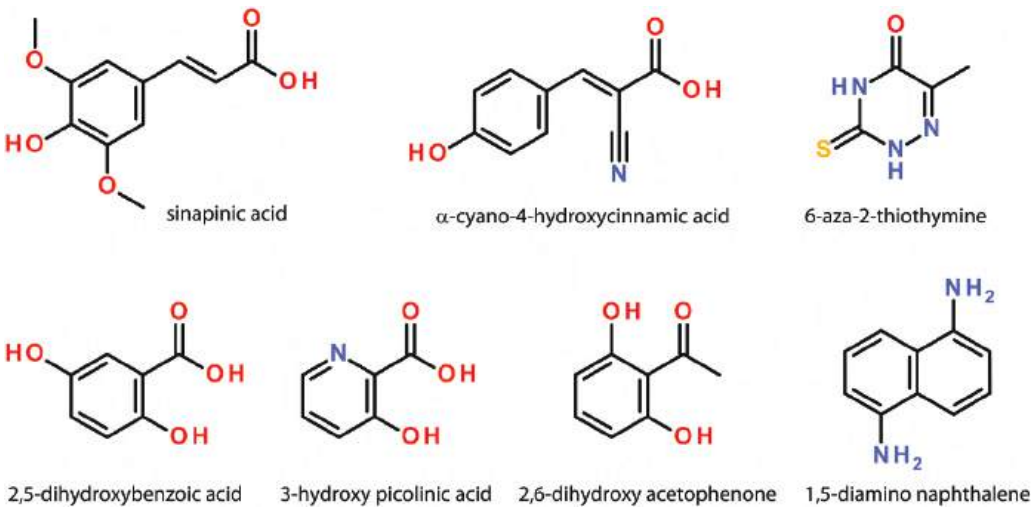


FIGURE 29.1: Examples of matrices used in MALDI ionization. The aromatic systems are excited at wavelengths of 266, 337 (nitrogen laser), or 355 nm. Water, acetonitrile, acetone, and chloroform plus 0.1% trifluoroacetic acid are popular solvents to dissolve the matrix. Many similar matrices to the ones shown exist that differ only by a few substituents. 1,5-diaminonaphthalene is not an acid but forms radicals upon laser irradiation that reduce disulfide bonds and cleave peptide bonds in proteins.

A pulsed UV laser, typically a nitrogen laser operating at 337 nm, excites and vaporizes the matrix. The matrix will transfer energy onto the macromolecule, thereby lifting it out of the solid state into the vacuum. Since the matrix is a weak acid it also helps to protonate the macromolecule, generating a positive ion. If the acidity of the matrix is not sufficient for protonation, small amounts of stronger acids such as 0.1% trifluoroacetic acid may be added. The number of protons transferred from the matrix onto the macromolecules is limited, and MALDI ionization yields ions with few charges. Peptide ions often carry only a single charge, whereas proteins are converted into a set of ions with different charges. In contrast, nucleic acids may retain their overall negative charge. Depending on the charge, the subsequent manipulation needs electric fields of different polarities. The different modes are termed *positive* and *negative ion modes* (Figure 29.2). The throughput of a MALDI measurement can be increased by using a sample plate holding 96, 384, or 1536 matrix-embedded samples that are ionized in sequence.

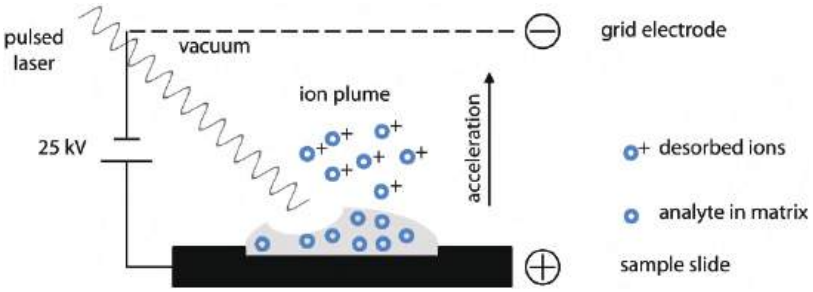


FIGURE 29.2: Sample preparation and ionization in MALDI. The macromolecule is embedded into a suitable matrix and placed under vacuum in the mass spectrometer. A laser pulse is applied to the matrix, which vaporizes both the matrix and the analyte into a plume. An electric field accelerates the ions according to their mass-to-charge ratio *m*/*z*. The depicted polarity corresponds to the positive ion mode.

29.1.1.2 Electrospray ionization

In ESI, the dissolved analyte is directly ionized from its solvent (Figure 29.3). The analyte is passed at a low flow rate of ~100 nL min^{-1} through a metal-coated glass capillary of ca. 100 µm diameter under atmospheric pressure. Special micro- and nano-ESI instruments use capillaries of down to 3–25 µm in diameter. The metal serves as the anode for a high voltage electric field of a few kV against a cathode in the direction of the detector. Charged droplets are ejected from the capillary by pressure and drawn into the spectrometer that is kept under vacuum. The droplets evaporate during this process. This is promoted by the vacuum, a carrier gas (usually nitrogen), and sometimes heating. With decreasing drop volume the charge density increases, up to a point where Coulomb repulsion between the ions in the droplets exceeds the surface tension of water that keeps the droplet intact, and the droplets separate into smaller and smaller droplets. At some point the solvent completely evaporates, leaving a varying number of charges on the surface of the ionized analyte (Figure 29.3). Proteins and peptides usually form positive ions, and similar to MALDI the protonation efficiency of the solvent can be increased by a trace of formic acid. Carbohydrates and nucleic acids form negative ions, and deprotonation can be promoted by addition of volatile amines or ammonia to the solvent. Like MALDI, ESI can be performed in positive and negative ion modes. In contrast to MALDI, more charges are transferred to the analyte, and the *m*/*z* distribution of the generated ions is also larger. Peptides usually carry a single charge after MALDI and 1–4 charges after ESI. For larger proteins, the charge differences are larger: 1–3 charges for MALDI and 20 to > 100 charges for ESI. As a rule of thumb, ESI of proteins produces ions that typically carry about one positive charge per five peptide bonds plus the charges on the basic amino acids Lys and Arg.

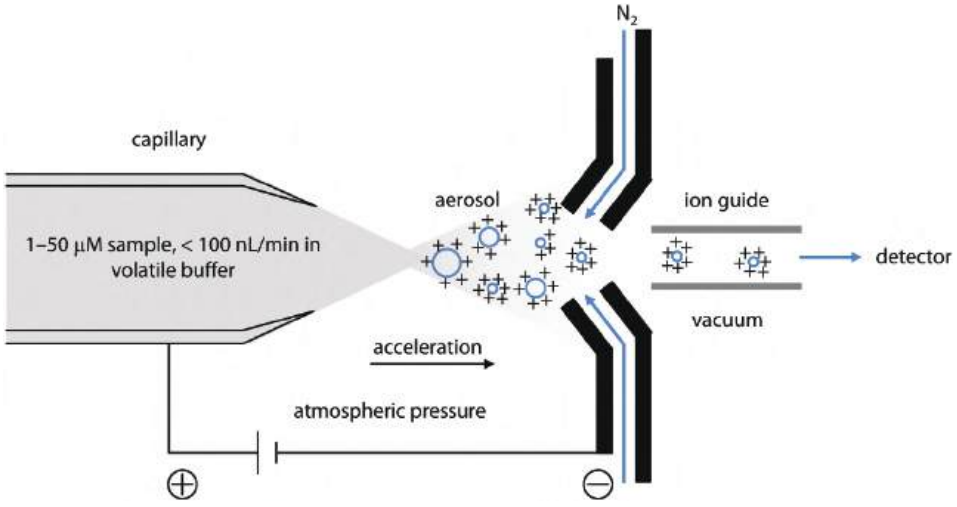


FIGURE 29.3: Principle of electrospray ionization. A sample in a volatile buffer is sprayed into an ionization chamber at atmospheric pressure where the solvent evaporates, aided by a stream of nitrogen gas. An electric field between the metal-coated quartz capillary and a cathode accelerates the ions, which are then guided into the evacuated mass spectrometer.

29.1.2 Ion storage and manipulation

The next step after ionization is acceleration of the ions in an electric field to select and separate them according to their m/z. Such devices are collectively called *mass analyzers*.

29.1.2.1 Time of flight analysis

One way to determine m/z is by measuring the time of flight (TOF) that an ion needs to travel through the electric field until it arrives at the detector. The simplest arrangement for TOF is an ionization chamber, an accelerating electric field, a field-free *drift tube* through which ions with small m/z pass more quickly than ions with large m/z, and a detector that measures the time-dependent impact of the ions. For this linear arrangement, the kinetic energy E_{kin} of the molecule after passing the electric field is given by

$$E_{kin} = zeU = \frac{1}{2}mv^2 \qquad \text{eq. 29.1}$$

where z is the number of elementary charges e, U is the voltage, m the mass of the ion, and v its velocity. We can rearrange eq. 29.1 to obtain m/z as

$$\frac{m}{z} = \frac{2eU}{v^2} \qquad \text{eq. 29.2}$$

The velocity is the distance L the ion travels through the drift tube, divided by the time t, the TOF measured by the detector. Substitution of v in eq. 29.1 by L/t gives an expression that allows calculating m/z from the measured TOF:

$$\frac{m}{z} = \frac{2eU}{L^2} \cdot t^2 \qquad \text{eq. 29.3}$$

According to eq. 29.3, m/z depends only on the TOF, and on instrument parameters that are constant and known in principle. Nevertheless, TOF mass analyzers are always calibrated to achieve higher reproducibility, either by running a sample of known molecular mass or by internal calibration, *i.e.* mixing analyte and reference with the matrix (for MALDI) or in the same buffer (for ESI).

The laser pulse in MALDI instruments generates a plume with a small spatial and time distribution of ions (Figure 29.2). Therefore, the kinetic energy after acceleration and hence the TOF along the drift tube vary for ions of the same m/z, and the spread of TOFs leads to lower resolution of the mass spectrum. The starting time of the acceleration can be synchronized when ion generation and acceleration are separated by switching off the accelerating field for a brief time interval after the laser pulse (*delayed extraction*). Ions leave the matrix with different kinetic energies. During the time between their generation and the acceleration, they now collide with molecules from the carrier gas, transferring some of their kinetic energy to the gas (Figure 29.3). Faster ions that have emerged with higher kinetic energy collide more often with gas molecules and lose more energy than slower ions, leading to a narrowing of the distribution of kinetic energies E_{kin}. In addition, faster ions have traveled further away from the anode toward the cathode by the time the accelerating field is switched on. As a consequence, they have a shorter distance to travel through the electric field, and they experience less acceleration than slower ions, which further narrows the distribution of kinetic energies.

The spread in E_{kin} can be further reduced by increasing the drift length for high-energy ions and reducing that for low-energy ions using a *reflectron* (Figure 29.4). A reflectron is an electric field of opposite polarity to the acceleration field placed at the end of the drift tube, serving as an ion mirror. Faster ions will continue to travel to the reflecting electrode for a longer time than slower ions with the same m/z. After reaching the reversal point, the faster ions are accelerated in the opposite direction for a longer time than the slower ions, entering a second drift tube. The net result is a focusing

of the ions at a certain point where the detector is placed. The total TOF is then a composite of the two drift times and the acceleration time in the reflectron.

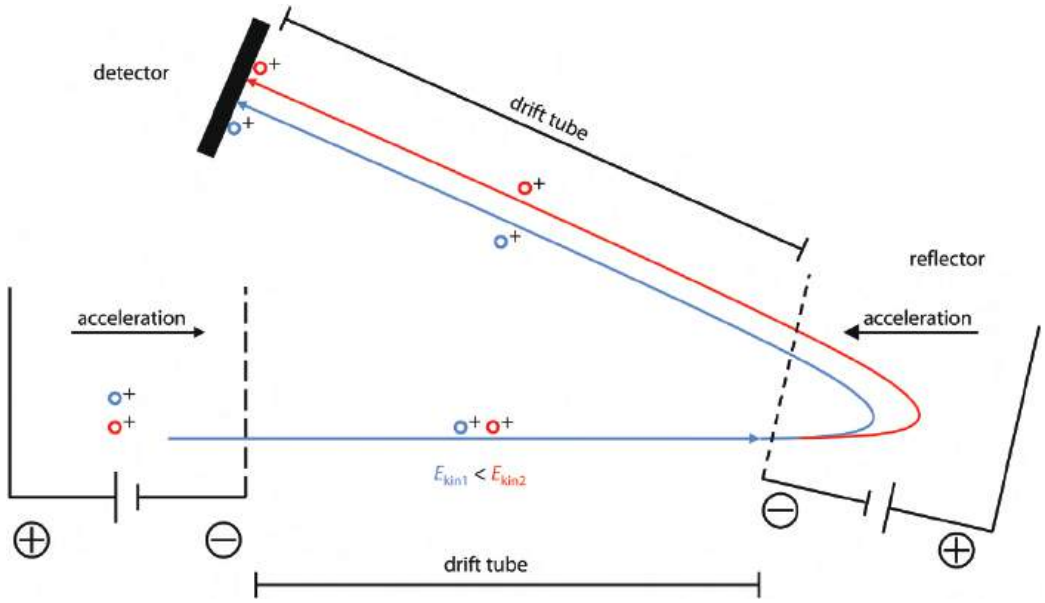


FIGURE 29.4: Harmonizing the TOF of ions with the same *m/z* by a reflectron. An ion with higher kinetic energy (red) travels deeper against the electric field into the reflectron than an ion with lower kinetic energy (blue). As a result, the ion with higher kinetic energy travels a longer distance (red trajectory) than the slower ion (blue trajectory). The overall effect is a harmonization of the TOFs for ions with different kinetic energies but equal *m/z*.

TOF analysis can be combined equally well with MALDI and ESI. Compared to MALDI, where a pulsed laser is used, and the ions are analyzed in the forward direction, the ionization in ESI is continuous, which means that ESI-TOF measurements need to be done in a direction perpendicular to the stream of droplets, not parallel to it. Other methods to analyze ions in the forward direction include quadrupole magnets and ion traps. Both of these can therefore be applied to MALDI and ESI.

29.1.2.2 Quadrupoles and ion traps

Quadrupoles and ion traps are used to select for ions with *m/z* ratios in a defined range. In a quadrupole, four rod-shaped metal electrodes are placed parallel to each other under superimposed alternate (AC) and constant (DC) voltages. Opposing pairs of rods have the same polarity and voltage. The constant voltage U_{DC} of a few kV serves to accelerate the ions between the electrodes: positively charged ions will oscillate between the positive electrodes and be attracted to the negative electrodes. The alternating voltage component U_{AC} of radiofrequency ω in the MHz range creates an alternating electric field that induces alternating deflections from the straight path along the quadrupole in the *x*- and *y*-directions. The amplitude of these deflections depends on *m/z*, the frequency ω, and the magnitudes of U_{AC} and U_{DC}. The combination of these parameters determines whether an ion may pass the quadrupole or has an unstable trajectory, either leaving the quadrupole or colliding with the electrodes (Figure 29.5). The quadrupole thus acts as a selective *mass analyzer* by only letting certain *m/z* ions pass. This may be understood by looking at the electrode pairs separately: positive ions with large *m/z* (heavy ions) are sent back and forth between the DC field of the positive rods but are not very much affected by the positive AC field because of their inertia. Thus, heavier ions pass the positive voltage filter of the quadrupole. Ions with too small *m/z* (light ions) are sensitive to the positive AC field: they are deflected too much and will crash into the electrodes or be ejected from the quadrupole, which puts a lower limit on the *m/z* range. A reverse argument

holds for the negative voltage pair of electrodes: ions with large *m*/*z* are attracted by the negative electrodes and will eventually crash into them or leave the quadrupole, unless they are "rescued" by a tuned negative AC field. Only lighter ions are sensitive enough to the radiofrequency field so that their trajectory is stabilized throughout the quadrupole, putting an upper limit on the *m*/*z* range. Together, the positive and the negative electrode pairs in the quadrupole create an *m*/*z* window that is adjusted by tuning the voltages and ω.

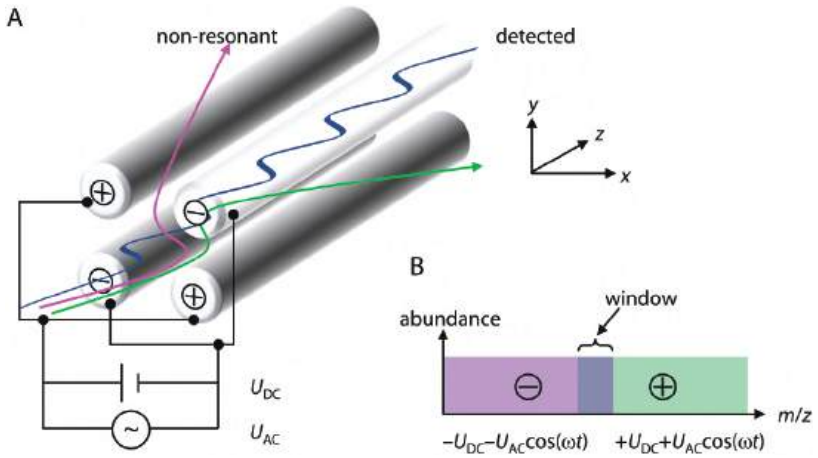


FIGURE 29.5: Schematic view of a quadrupole. A: One of the electrodes is rendered transparent to provide a view onto the path of the filtered *m*/*z* ions that are passed on to the detector. The negative electrodes are connected and have the same voltage. The same is true for the positive electrodes. Only ions with a certain *m*/*z* undergo stable oscillation while traveling through the linear quadrupole and can pass to the detector (blue trajectory). Non-resonant ions (magenta and green trajectories) will leave the quadrupole or crash into the electrodes. B: Only low *m*/*z* ions pass between the negative rods, and only high *m*/*z* ions pass between the positive rods. Those ions that meet both conditions form a window of detection, which is scanned by changing the frequency ω and the AC and DC voltages.

Quadrupoles act on moving ion beams, as do reflectrons in TOF analysis. Ion traps, on the other hand, not only select ions with certain *m*/*z*, but in addition have cleverly shaped electric fields that can decelerate and store ions for subsequent analysis. An electrostatic ion gate controls entry and exit of ions, and thus the number of ions stored in the trap. The trap contains a small amount of a collision gas (ca. 0.1 Pa of He). Upon entry into the trap, the velocity of ions is reduced by collisions with the gas molecules. The ions are then focused by an electromagnetic lensing system, a three-dimensional circular electric field that is created by a ring-shaped electrode operating with voltage U_1 and a constant radiofrequency ω of typically 1.1 MHz (Figure 29.6).

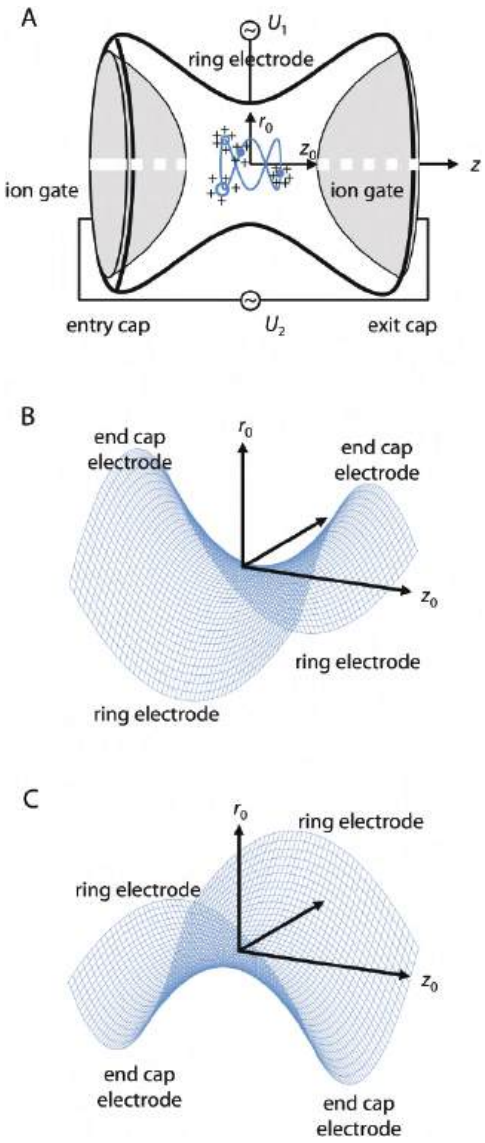


FIGURE 29.6: Schematic of an ion trap. A: Overview of the trap. Ions of a certain *m/z* range, defined by the amplitude and frequency of the alternating voltage at the ring and cap electrodes, U_1 and U_2, are kept on stable orbits in the trap. Varying the parameters of U_2 permits fragmentation and/or ejection of ions with a defined *m/z* for further analysis. The holes in the entry and exit caps are opened and closed by an electrostatic ion gate. The trap geometry is defined by the curvatures of the ring electrode and the caps, and by the radii r_0 and z_0. B: The shape of the electromagnetic field in the trap is shown by a blue mesh. Ions are focused toward the saddle point in the *z*-direction but may diffuse perpendicular to the *rz*-plane. C: The field spanned by the ring and cap electrodes is inverted to re-focus the ions toward the saddle point in the *rz*-plane. Oscillation of the field forces the ions into a stable orbit.

The electric field U at the position (r,z) generated by the ring electrode can be described by eq. 29.4.

$$U(r,z) = A \cdot \left(z^2 - \frac{r^2}{2} \right) \qquad \text{eq. 29.4}$$

A is a constant related to trap geometry. The electric field has the shape of a saddle with a minimum at the center of the ion trap. The saddle focuses the ions in direction r but they can continue to move in the z-direction. This defocusing is counterbalanced by switching the polarity of the ring and end

cap electrodes, which flips the radiofrequency field: the ions now focus back in the *z*-direction, but can defocus in *r*. The balance between these field directions keeps ions of a certain *m*/*z* orbiting in the trap. At the same time, the ions are bunched up in the center of the trap, which is useful for pulsed ejection into a mass analyzer such as the orbitrap (Section 29.1.2.3). The balance of the fields is key for the performance of the ion trap, which is why the number of ions within the trap is controlled: too many ions result in *space charge* effects that distort the electric fields of the trap, which in turn reduces its performance.

The selectivity of the ion trap, *i.e.* the *m*/*z* of ions that have stable orbits, is determined by the geometry of the trap, the magnitude of U_1, and the radiofrequency ω. While a quadrupole is usually tuned to allow passage of a narrow *m*/*z* range, an ion trap can hold a wider range of *m*/*z*.

The ejection of ions from the trap in a controlled manner is achieved by applying an oscillating electric field U_2 to the end caps of the trap. Depending on the magnitude of U_2, the orbit of ions with a narrow range of *m*/*z* becomes unstable. In this *mass instability mode*, the ions are in resonance with the oscillating electric field U_2 and ejected from the trap along the *z*-axis. Quadrupoles and ion traps can be placed in sequence for finer sorting and more accurate manipulation of ions, an arrangement referred to as *tandem*-MS, or MS/MS.

29.1.2.3 Orbitraps

The orbitrap is a type of mass analyzer that is frequently used in modern MS equipment for proteomics and native mass spectrometry. Like ion traps, orbitraps serve to select, store, and manipulate ions of a certain *m*/*z* range. Orbitraps can hold many ions at once and of very different *m*/*z* (Figure 29.7). In contrast to ion traps, orbitraps combine mass analysis with ion detection and *m*/*z* determination.

The orbitrap consists of an inner spindle-shaped electrode that is co-axially surrounded by a cylindrical outer electrode, and two end cap electrodes. A static DC voltage is applied between the inner and outer electrodes, and an alternating AC voltage between the end caps. The DC voltage produces a radial logarithmic potential while the AC voltage produces an axial quadrupole potential, which together establish a "quadro-logarithmic" potential (eq. 29.5):

$$U(r,z) = A \cdot \left(z^2 - \frac{r^2}{2} + B \cdot \ln r \right) \qquad \text{eq. 29.5}$$

The constants *A* and *B* in eq. 29.5 are specific to the geometry and voltages of the orbitrap, and determine the shape of the electric field. Ion bunches are injected tangentially to the inner electrode and immediately begin to orbit the electrode due to the Lorentz force (Figure 29.7). The centrifugal force of the ions counteracts the Lorentz force until an equilibrium radius is reached. Ions of different *m*/*z* have different velocities and therefore different rotational frequencies and orbiting radii. The radii can be reduced to a desired minimum value by increasing the voltage to 3.5–5 kV, which in turn increases the detectable mass range because ions with larger *m*/*z* can now be stored in the orbitrap. At the final voltage of the orbitrap, those ions with the same *m*/*z* organize into rings.

Importantly, by injecting the ions away from the center (the thickest part) of the inner electrode at $z \neq 0$, they experience an additional lateral (axial) movement along the *z*-direction because of the quadrupole field. The special shape of the orbitrap electrodes ensures that the axial oscillation along the *z*-direction is harmonic, meaning that it is independent of the perpendicular radial oscillation around *z*. The angular frequency ω of this axial oscillation along *z* is

$$\omega = \sqrt{\frac{z}{m} \cdot k} \qquad \text{eq. 29.6}$$

where *k* is the force constant of the potential, similar to a spring constant (Sections 17.2.1.2 and 21.4.1). In fact, the axial oscillation solely depends on the *m*/*z* of the ions, and it is this axial oscillation that is measured in an orbitrap.

Once the voltage of the orbitrap has reached its final value, the electromagnetic field becomes static and detection can start. The trapped ion rings induce a current in the outer electrode, the *image current* (Figure 29.7). The current is measured over a few milliseconds to several seconds and Fourier transformed (Section 31.15) from the time domain into the frequency domain to give the frequencies ω of the ions. The *m*/*z* ratios for all ions are then calculated using eq. 29.6. Thus, all ions are detected simultaneously over a given time. The resolution of the orbitrap is increased both by increased observation time and by increased field strength.

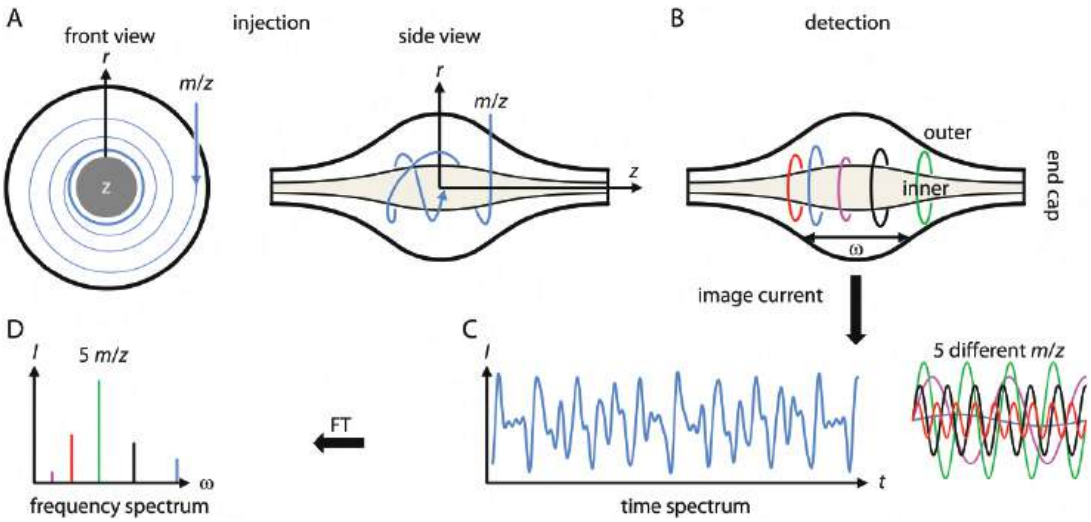


FIGURE 29.7: Ions in an orbitrap. A: Orbitraps have specially shaped inner (gray fill) and outer electrodes to which a constant DC voltage of 3.5–5 kV is applied. The radial distance *r* between the electrodes varies as a function of the axial coordinate *z*. Ions are injected in bunches from another ion source (blue arrow in the front view), and oscillate radially around the cathode until centrifugal and electrostatic forces are in equilibrium. By injecting the ions away from the center of the inner electrode (side view, $z \neq 0$), an oscillation in the *z*-direction is started that is independent of the oscillation in the *r*-direction. The side view shows an example of a stable ion trajectory, which resembles a helix. B: The ions on stable orbits organize into rings that oscillate laterally with frequency ω. C: For analysis, the image current at constant voltage is observed over a period of milliseconds to several seconds. Here, the image current is the sum of five sine waves of different amplitudes and frequencies shown on the right-hand side. D: Fourier transformation (FT) of the time spectrum into the frequency spectrum yields the *m*/*z* values of all ions simultaneously according to eq. 29.6.

29.1.2.4 Ion fragmentation and sequencing

In the analysis of macromolecular ions it is often desirable to fragment them into smaller ions with *m*/*z* ratios that can be handled with greater accuracy and measured to higher precision. Sequencing of proteins and in principle also nucleic acids by mass spectrometry relies on formation of a series of ions that differ by one amino acid or one nucleotide. The general strategy for sequencing is to first collect a complete (and very crowded) mass spectrum of a sample that has been fragmented and then use mass analyzers to isolate a certain fragment from the sample by its *m*/*z* that is then analyzed in a second MS run.

Ion fragmentation involves breaking of chemical bonds (Figure 29.8). Conditions for fragmentation need to be controlled to ensure reproducible fragmentation and to minimize side-reactions to different ions that complicate the spectra obtained. Fragmentation of macromolecular ions is achieved by electron bombardment (not selective, but useful for small molecules), light irradiation (in MALDI), and collision of ions with other ions and with carrier gas molecules in drift tubes and ion traps.

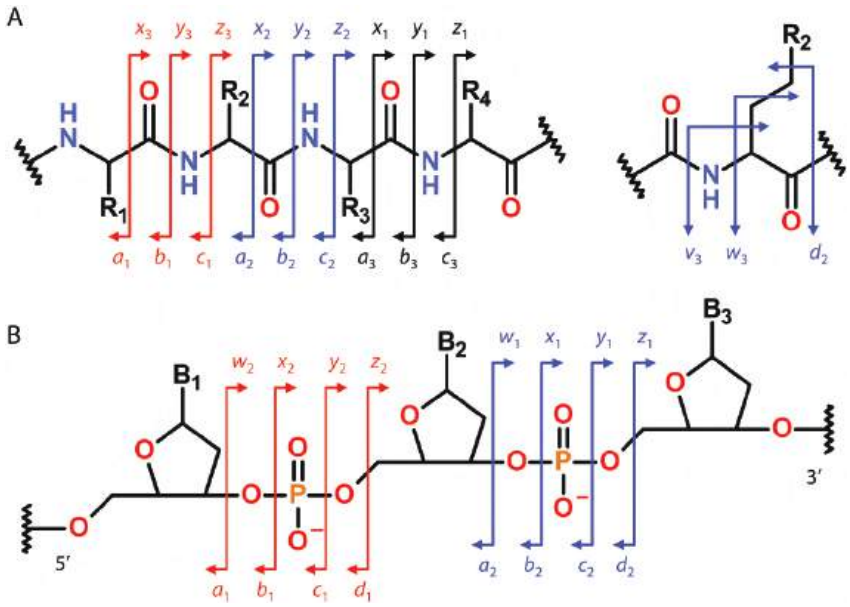


FIGURE 29.8: Backbone fragmentation of peptides and nucleic acids. A: Bond breakage in the peptide backbone yields the N-terminal *a*, *b*, and *c* series and the C-terminal *x*, *y*, and *z* series of fragments. Breaking of the peptide bond leads to the *b*/*y* fragments (Figure 29.9). Certain side chains can also break during fragmentation (top right), leading to the *d*, *v*, and *w* series. B: In nucleic acids, the bonds around the phosphodiester preferentially break, giving the 5′-series *a*, *b*, *c*, *d* and the 3′-series *w*, *x*, *y*, *z*. The subscripts refer to the position of the amino acid or nucleotide. Usually one fragment is neutral and cannot be detected. If two or more bond cleavage events occur within a single peptide, internal fragments are produced.

The first possible source of ion fragments is MALDI when the laser light is of such high intensity to not only ionize the macromolecule to a *parent ion*, but to immediately produce *daughter ions* by direct light absorption or temperature effects that lead to bond fission. Further ions are generated in MALDI by collisions between ions within the desorption plume shortly (ns–µs) after excitation of the sample by the laser pulse, while the ions are accelerated by the electric field (Figure 29.2). This mode of ion fragmentation is termed *in-source activation* or *in-source decay* (ISD). Some ions produced by MALDI are metastable and spontaneously decay while traveling along the drift tube, which is termed *post-source decay* (PSD). A third method often used in modern mass spectrometers to produce ion fragments is *collision-induced decay* (CID). Drift tubes or ion traps are filled with a *carrier gas* or *damping gas*. Collisions between ions and gas molecules lead to ion fragmentation. For ion traps, the procedure is as follows: in a first MS run, a subset of ions with the same m/z is selected by discarding all other ions. Next, the selected ions are fragmented directly within the ion trap. Instead of ejecting an ion by application of a large alternating voltage U_2 (Figure 29.6), a slightly smaller voltage, the *tickle voltage*, is applied to the end caps. The ions are accelerated and collide more and more frequently with the carrier gas. When the kinetic energy of the ions is sufficient, collisions lead to bond fission and generation of daughter ions. Only then are the daughter ions ejected from the ion trap and directed toward the detector in a second MS. Such concatenation of several mass analyzing steps is another example of tandem-*MS* or MS/MS.

Peptides fragment mainly at the peptide bond (Figure 29.9). The other two backbone bonds in the polypeptide may also be broken, but these processes are less efficient than breakage of an amide because they involve the more stable C-C bonds.

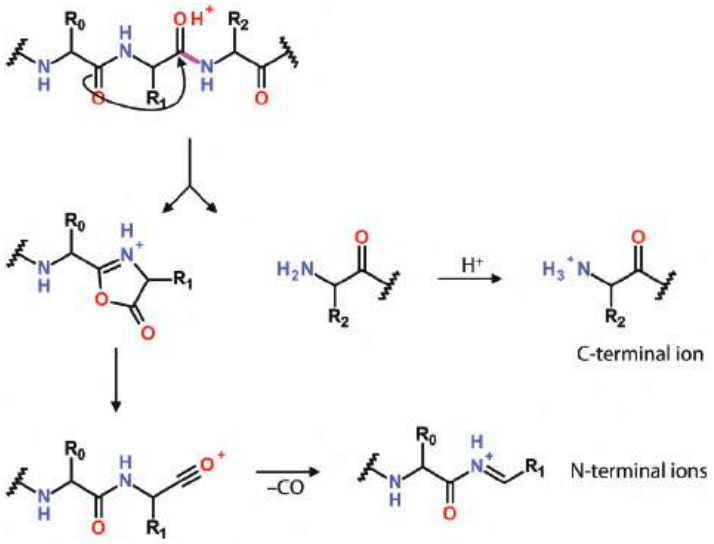


FIGURE 29.9: Main products from cleavage of peptide ions. Protonated carbonyl groups are electrophilic, and can be attacked by the carbonyl group of a preceding amino acid. The amino group of the C-terminal part attracts a proton to form the *C-terminal ion* (*y*-ion). The N-terminal intermediate ring can open to form an oxonium ion, which may or may not eliminate carbon monoxide before detection as an *N-terminal ion* (*b*-ion). Usually, one part of the peptide remains neutral and is not detected.

For a singly charged parent ion, the localization of the positive charge determines whether the N- or C-terminal daughter ion is detectable, the neutral fragment is not detected. If the parent ion has two or more charges, both daughter ions may be charged and are detected. The fragmentation reaction thus produces a series of daughter ions that differ by a single amino acid residue. The identity of this amino acid is inferred from the mass difference between two successive daughter ions. From the series of daughter ions derived from a peptide fragmented at different peptide bonds, the peptide amino acid sequence is deduced (*de novo* sequencing). Although normally each peptide only undergoes a single fragmentation reaction, sometimes fragmentation occurs at two sites, generating internal ions and more complex mass spectra. Ions that undergo fragmentation through reactions at their side chains (Figure 29.8) can also complicate the interpretation of mass spectra.

29.1.3 Detection

We have now treated the generation of ions and their analysis and separation according to the m/z ratio. All of these ions eventually need to be detected to generate mass spectra. Detector types include ion-to-photon detectors, electron multipliers, and microchannel plates. In ion-to-photon detectors, ion collisions with a scintillating compound generate photons that are then detected as a photocurrent by a photomultiplier tube. In *electron multipliers* (Figure 29.10), the ions impinge on an electrode, the *conversion dynode*, and liberate primary electrons. This initial signal is too small to be detected accurately, and is therefore amplified by acceleration of the primary electrons in an electric field toward another dynode, where more electrons are liberated. A cascade of dynodes leads to an avalanche of electrons that are detected at the last electrode, the *detection anode*.



FIGURE 29.10: Principle of a discrete electron multiplier. An ion collides with the conversion dynode, liberating a few secondary electrons. Attraction of these electrons by a sequence of dynodes produces more and more electrons until a gain of typically 10^6 is reached.

Microchannel plate detectors (MCPs) are particularly sensitive and versatile arrays of dynodes (Figure 29.11). About ten million glass tubes of ~6–10 µm diameters and 1 mm length, coated with a semi-conductor on the inside, are densely packed into a plate. The tubes or channels are parallel to each other, but often inclined by ca. 8° from the plate normal to ensure that any electron (from a conversion dynode) entering from the cathode side collides with the inside wall of the channel. The collision with the semi-conducting walls liberates more electrons, which continue to collide with the walls and generate more secondary electrons while passing along the channel. At the anode side of the MCP, a cloud of several thousand electrons exits over a narrow area, and is detected as a signal current. The large area of MCPs offers high spatial resolution. Since only a few channels out of several millions are used for the detection of a certain m/z, many different ions hitting the MCP at different spots can be detected simultaneously. MCPs have fast time responses of a few nanoseconds to below one nanosecond, which makes them especially useful for TOF measurements because they do not spread the TOF signal appreciably.

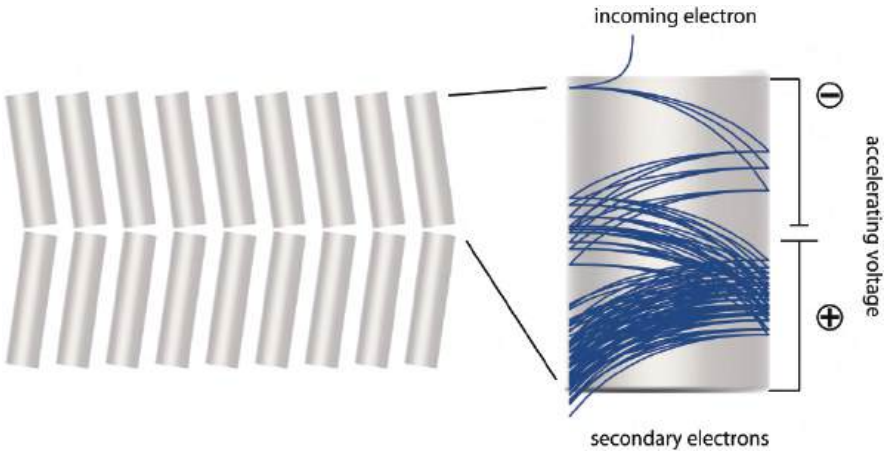


FIGURE 29.11: Schematic principle of an MCP. Two plates are stacked on top of each other to limit ion feedback: any signal going into the opposite direction is blocked at the interface of the two plates. The individual micro channels have a slight inclination from the normal. The right-hand figure visualizes how electrons are multiplied in a single tube for detection upon exit at the bottom.

29.1.4 Mass spectra

The collection of peaks corresponding to ions with different *m*/*z* ratios that is detected after ionization, manipulation of ions, and mass analysis is the mass spectrum. MALDI generates only a few ions from the molecule analyzed, with a single or a few charges. ESI generates a series of ions with multiple charges and hence different *m*/*z* ratios from each molecule in the sample analyzed, giving rise to a series of peaks. As a consequence, MALDI spectra with only a few peaks can be analyzed directly, whereas ESI spectra are more complex (Figure 29.12).

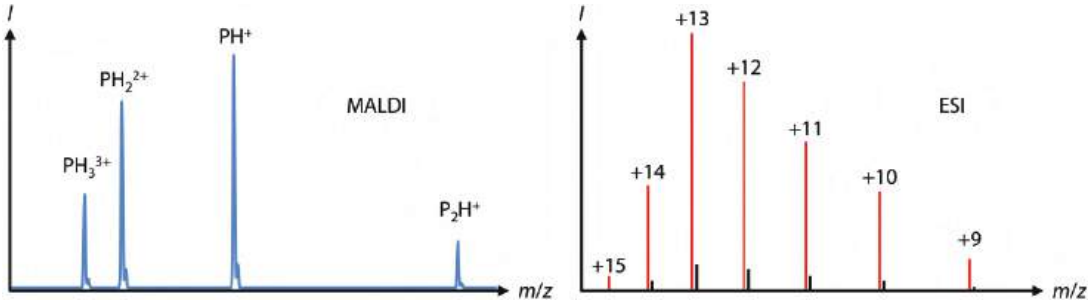


FIGURE 29.12: Schematic MALDI and ESI spectra of a peptide. For MALDI, the most prominent ion is the singly charged molecular ion. Ions with two or three charges are less abundant. A small population of ions P_2H^+ with protein dimer is also sometimes detectable. Ions from ESI contain many more charges per molecule than from MALDI. A smaller series of adduct peaks is visible to the right of each *m*/*z* peak.

The *m*/*z* ratio for each peak in a mass spectrum follows the relation

$$\frac{m}{z} = \frac{M_M + n\left(H^+\right)}{z} \qquad \text{eq. 29.7}$$

where M_M is the molecular mass, $n(H^+)$ the number of protons attached, and z the number of positive charges. The smaller number of peaks in a MALDI spectrum can often be interpreted directly. In addition to molecule ions with multiple charges, non-covalent adducts of ions such as sodium, potassium, matrix ions, and phosphate may be observed. The identity of these adducts can be determined from the mass difference to the unmodified ion. For instance, in sodium adducts, Na^+ (atomic mass 23) replaces a proton without changing z, which results in a +22 Da peak (Figure 29.12). Similarly, matrix adducts are distinguished from the parent ion by tabulated molecular mass shifts.

The more complex ESI spectra typically require deconvolution to obtain the molecular mass of the detected ions. Initially, the absolute number of protons n (eq. 29.7) is not known, but must be an integer. To determine the molecular mass M_M of the molecule that gives rise to a series of peaks, n is guessed for one of the detected peaks, and M_M is calculated according to eq. 29.7, taking into account that ions in neighboring peaks must differ by one charge. The value for n which gives identical masses M_M for all peaks of the series is the correct number of charges, and M_M is the correct molecular mass of the molecule analyzed. Ion adducts give rise to smaller peaks on the right-hand side of the main *m*/*z* peaks in an ESI spectrum. They occur at the same *m*/*z* difference to each main peak, and can be identified from this mass difference.

29.1.5 Applications

29.1.5.1 Mass analysis for the identification of molecules

Determination of the molecular mass of the molecule analyzed is the main purpose of MS experiments. Mass determination on its own allows the nature of a compound, its isotope distribution, and its purity to be established. Mass determination by MS achieves accuracies of below 0.01%. In other words, the calculated and the measured molecular masses for a macromolecule of 10 kDa can be compared to an accuracy of 1 Da. Therefore it is possible to distinguish reduced from oxidized forms

of cysteines (2 Da difference), which enables the number of disulfide bonds in small proteins to be counted. Preparations of macromolecules are often heterogeneous due to chemical modifications. If a mass difference between expected and measured molecular masses is found, databases can be searched to tentatively assign the discrepancy in mass to a type of chemical modification. Post-translational modifications of proteins (Section 18.1.4) can also be identified by MS. While defined modifications such as phosphorylation or methylation are readily identified, other modifications may be heterogeneous and the resulting complex mass spectra may be difficult to assign. For example, proteins may be glycosylated at Asn, Ser, or Thr residues. Depending on the source of the protein, the glycosylation trees can be heterogeneous. In such cases, identification is not possible without fragmentation and analysis of the glycosylated fragments. MS is also routinely used to identify products in crosslinking experiments of macromolecular complexes from their mass, and to map binding interfaces by identification of the crosslinking sites (Section 29.1.5.6).

29.1.5.2 Isotope distribution and isotope exchange

The resolution of MS is sufficient to detect differences in isotopic composition, for example in the content of ^{14}N and ^{15}N, or ^{12}C and ^{13}C. Isotope analysis is helpful to test the percentage incorporation of ^{13}C or ^{15}N of pure NMR samples. The natural abundance of ^{13}C is about 1.1% (the rest is mainly ^{12}C, and traces of radioactive ^{14}C; see Box 7.2). The probability to contain more than one ^{13}C increases with the number of carbon atoms in the molecule. For organic molecules or small peptides, the likelihood of incorporation of more than one ^{13}C is small. For proteins, the likelihood of incorporation of several ^{13}C nuclei increases with size. The distribution of peak intensities within a series reflects the probability for the particular isotope content (Figure 29.13). The shape of the peak distribution is a skewed Gaussian distribution, and the relative peak intensities can be calculated from the isotope abundances using the binomial theorem (Section 31.3).

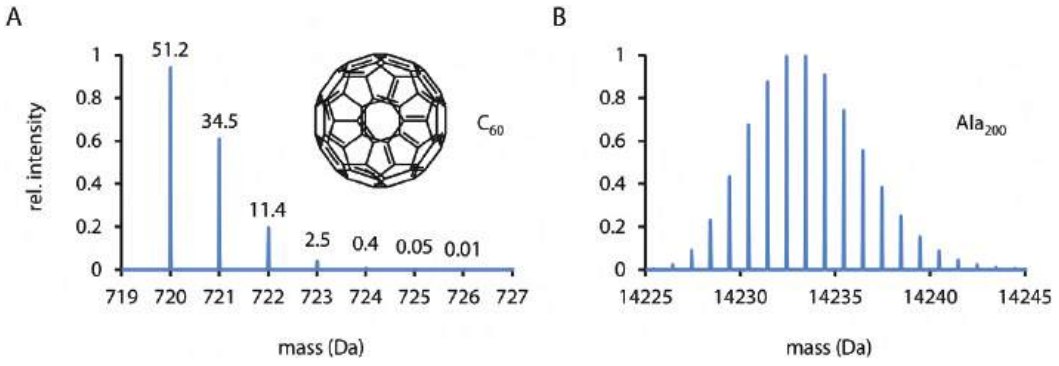


FIGURE 29.13: Skewed Gaussian peak distribution in small molecule and protein ESI spectra. All peaks have $z = 1$. A: Small organic molecules such as the fullerene C_{60} and short peptides have a quasi-exponential distribution of intensities for their m/z peaks. B: Isotope distribution of the hypothetical protein poly-Ala_{200} of $M_M = 14233.6$ Da shows a skewed Gaussian distribution of peak intensities.

For mixtures of macromolecules, the natural isotope distributions can lead to very complex mass spectra. For example, a 100 kDa protein contains ca. 50 ^{13}C atoms, leading to at least 50 different peaks in the mass spectrum for the same molecule. Two macromolecules of similar masses will have overlapping mass spectra. Some of the peaks cannot be resolved, which limits the resolution attainable from mixtures.

Isotope exchange experiments, particularly hydrogen-deuterium (H/D) exchange, are frequently used to probe the accessibility of hydrogens in proteins or other macromolecules. Deuterium (^2H or D) has an extra neutron and thus twice the mass of ^1H. Under mild alkaline conditions, hydrogens in peptide bonds and any side chain that carries a NH or OH group can exchange with deuterium from the solvent. When a protein is placed in an aqueous buffer based on D_2O instead of H_2O, the velocity of H/D exchange depends on the hydrogen accessibility: in solvent-exposed regions, hydrogens exchange quickly, while hydrogens in buried parts exchange slowly or not at all. After

stopping the reaction by lowering the pH, the protein is fragmented, and the peptides are analyzed for deuterium content (and sequence) by MS. Peptides with high deuterium content reflect parts of the protein that are solvent-exposed. H/D exchange experiments can also be performed in a time-dependent manner by stopping the exchange reactions at different time intervals. Time-resolved H/D exchange is frequently employed to study protein (un)folding mechanisms (Section 18.5.5.3). Also, binding partners protect their binding sites from hydrogen exchange because they exclude access of solvent. H/D exchange monitored by MS can thus be used to follow complex formation. Changes in protection patterns can also give insight into conformational changes.

29.1.5.3 Protein identification from one- and two-dimensional gels

MS is routinely used to identify individual proteins separated by one- or two-dimensional gel electrophoresis. A band or spot containing the protein of interest is excised from the gel, and the protein is hydrolyzed *in situ* by a specific protease (*in-gel digest*), *e.g.* trypsin. The resulting mixture of peptides is first separated by liquid chromatography (LC) before determining their masses by MS. LC-MS couples the chromatographic separation directly to the ionization chamber of an ESI mass spectrometer. If the peptides are too large, further fragmentation of the proteolytic peptides in the mass spectrometer is done by CID or PSD (Section 29.1.2.4). The peptide masses from the unknown protein are compared to calculated masses (using the same protease *in silico*) of all proteins in a sequence database. In favorable cases, the unknown protein can be identified based solely on the masses of its proteolytic peptides, without any sequence information. At least two unique peptide masses are required to identify the unknown protein. For a peptide mass hit to be statistically significant, *i.e.* to not occur by chance, a low *Expect value (E-value)* is required. The *E*-value describes the number of hits for a certain peptide mass we expect to get purely by chance when searching a database of a particular size for that mass: the *E*-value represents the random noise of the database. The larger the peptide, the more statistically significant will a hit from this peptide be in the database, because the likelihood of finding the same peptide mass twice or more times in the database is smaller for longer sequences.

Protein identification by MS becomes more difficult when post-translational modifications such as glycosylation and disulfide bonds are present, or when side-chain fragmentation increases the complexity of the mass spectrum. However, the bioinformatics tools to calculate expected fragmentation patterns from the sequence databases are being constantly expanded to include masses for such unusual peptides. Hence the identification of proteins from peptide masses becomes more and more facile. However, if the mass-based approach does not work, some representative peptides are sequenced by MS, and run against the sequence database using the additional sequence information.

29.1.5.4 Native mass spectrometry

The crystalline matrices, acids, and organic solvents normally used in MALDI or ESI will dissociate non-covalent macromolecular complexes and unfold proteins. While MALDI is usually incompatible with maintaining the native state, ESI can be performed on native proteins and macromolecular complexes by replacing the organic solvent with a volatile buffer such as ammonium acetate, which in vacuum decomposes into NH_3 and acetic acid, leaving the molecules in their native state. Despite the more complex spectra obtained by ESI (more peaks than MALDI) and the associated difficulties in their interpretation (overlap of peaks), ESI is the preferred method of ionization for native MS.

Proteins are stabilized to a large extent by the hydrophobic effect (Section 18.5.1). The hydrophobic effect vanishes in the vacuum of the mass spectrometer, where the dielectric constant ε_r is unity. It was initially believed that proteins would instantaneously unfold and complexes held together by hydrophobic interactions would immediately dissociate *in vacuo*. Nevertheless, it is now firmly established that native protein conformations can prevail in a mass spectrometer, at least on the millisecond time scale of data acquisition. There are several reasons why proteins and their complexes remain folded and intact *in vacuo*: (1) attractive van der Waals interactions, (2) increased hydrogen bonding of surface residues with each other when bulk water is absent, and (3) stronger charge-charge interactions between ionized residues in complex interfaces. These arguments also

hold for membrane proteins in the absence of detergent molecules. It is now possible to ionize membrane proteins and strip them of detergents without triggering their unfolding.

Advances in instrumentation and ionization procedures have made large molecular complexes amenable to native MS studies. In general, the larger the molecular mass, the more adducts (H^+ and other cations) are formed, and also the broader the isotope distributions. The higher complexity of ESI spectra limits the resolution with which masses can be determined from native samples of high molecular mass. However, complexes isolated in their native form from cell extracts or from membrane preparations are analyzed by native MS for their conformational distributions (by ion mobility; Section 29.1.5.5), their stoichiometry (by quantitative MS; Section 29.1.5.8), their subunit interactions, and their assembly pathway, which is not possible with denaturing ESI.

Subunit interactions within a large complex are studied by generating sub-complexes with an overlapping set of subunits in solution. The decomposition of these protein sub-complexes *in vacuo* follows a series of well-defined steps: first, water molecules and counter ions are removed by collision with gas molecules followed by structural changes of the protein complex. The next steps are unfolding and dissociation of single protein subunits from the complex, one after the other. The two masses of the liberated monomer and the remaining complex are detected separately, establishing which subunit has dissociated from the complex. In case different decomposition pathways occur at the same time, more than one pair of liberated monomer and remaining complex is produced. At the end, a complete assembly chart for the complex may be constructed.

In analogy to H/D exchange (Section 29.1.5.2), whole subunits can be exchanged in a multi-protein complex, and the exchange kinetics can be monitored. For these experiments, the mass of the subunits may either be known or they need some sort of mass tag, such as an isotopic label or a chemical modification. Exchange experiments with several subunits of the complex can thus reveal the assembly pathway. Protein-nucleic acid complexes and complexes of membrane proteins with soluble proteins can also be studied by native MS. Examples of membrane proteins studied are the rotary V-ATPase and the P-glycoprotein. Large multi-protein/nucleic acid complexes that have been studied by native MS are the ribosome, the TRAP (tryptophan regulator attenuation protein), and the spliceosome.

29.1.5.5 Ion mobility and molecular shape

In contrast to reflectrons, quadrupoles, and ion traps, which all select ions based on m/z, ion mobility measurements separate ions that have the same m/z ratio but differ in shape. Different mobilities of ions of the same m/z are caused by different conformations of the macromolecule. Ions with extended conformations have large *cross-sections* while more compact ions have smaller cross-sections. To resolve ions by shape, they are passed through a drift tube where they are slowed down by collisions with a carrier gas. Ions with large cross-sections experience more collisions with gas molecules compared to compact ions. The resulting distribution of velocities and the resulting distribution of drift times, measured by TOF, allow conclusions to be made on the conformational space that is sampled by the protein. Ligand-induced conformational changes can also be detected by ion mobility. The drift times can be converted to orientationally averaged collision cross-sections, which represent the cross-sections of all possible orientations of a particular macromolecule or complex. If structural information such as a crystal structure or an envelope from EM or SAXS is available, the cross-section calculated from the structures can be compared with that from MS to find which conformation is present *in vacuo*.

29.1.5.6 Identifying protein-RNA interaction sites after photo-crosslinking

RNA-binding proteins (RNPs) in the cell are involved in vital processes such as transcription, splicing, and translation. To identify the interaction sites between proteins and RNA molecules in the complex, they can be subjected to photo-crosslinking, followed by MS analysis. Photo-crosslinking of RNA-protein complexes can be done in the cell or in solution by irradiation with UV light at 260 nm, where the nucleobases absorb (Section 21.2.6) and crosslink to nearby proteins. Alternatively, 4-thiouridine can be incorporated into the RNA, and crosslinking is performed by irradiation at 365 nm.

The complexes are isolated from the cells and enriched by chromatography. Instead of analyzing the entire protein-RNA complex by MS, the analysis is simplified by trimming the complex with RNase and proteases to generate fragments of only 1–3 nucleotides that are crosslinked to a short peptide. The masses of the thousands of peptides and crosslinked RNA-protein adducts are then determined by LC-MS. To find the masses corresponding only to the adducts, an identical control experiment with a non-crosslinked sample is run in parallel. The identified RNA-protein species are then sequenced by MS/MS to identify both the RNA and protein sequences that are in contact in the complex.

In an experiment that crosslinked proteins and RNA in yeast ribosomes, from almost 10000 determined masses only 184 (ca. 2%) remained as candidates for crosslinked species. Their identification provided the crosslinking sites and thus information on the RNA-protein interfaces within the ribosome. The analysis also showed that proteins can crosslink to RNA through almost all amino acids except D, N, E, and Q, whereas RNA forms crosslinks almost exclusively through uridine or 4-thiouridine (if it was included during RNA synthesis). While most of the identified RNPs were ribosomal proteins, some unexpected enzymes including enolase and inorganic pyrophosphatase were identified as RNA-binding proteins. These proteins might have ribosome-related functions that were previously unknown.

29.1.5.7 Secondary ion mass spectrometry

Secondary ion mass spectrometry (SIMS) is a MS-based method for label-free two- and three-dimensional imaging of biochemical surfaces such as cell membranes. The surface is bombarded with heavy *primary* ions with energies of tenths of a keV that are focused on sub-µm spots. The primary ions generate *secondary* ions of 1 kDa mass or less from lipids and amino acids, reporting on the distribution of membranes and proteins, respectively. The secondary ions are detected in parallel, yielding a full mass spectrum for each spot on the surface interrogated. From the mass distributions and abundances of the secondary ions, the chemical composition of the surface is inferred as a function of position in two dimensions.

The technique was applied to monitor the change in lipid composition of the *Tetrahymena* cell membrane during mating. *Tetrahymena* is a common fresh-water protozoon that forms curved fusion pores at the conjugation site during mating. For SIMS analysis, the cells before and during mating were freeze-fractured and the lipid composition was determined using an In^+ primary ion beam. Analysis of the secondary ions by MS revealed a reduction of the phosphocholine signal at m/z = 184, matched by an increased signal for the head group of 2-aminoethylphosphonolipids at m/z = 126, suggesting a dramatic change in lipid composition during mating.

Three-dimensional information with sub-cellular resolution can be gained by a series of two-dimensional scans that are stacked on top of each other. However, because bombardment of surfaces with primary ions partly destroys sub-surface structures, a fresh layer of intact material needs to be exposed by a gentle ablation step prior to the next two-dimensional scan. The three-dimensional distribution of triacylglycerides in *Xenopus* oocytes at ambient temperature was analyzed this way by cycles of two-dimensional imaging of the surface by SIMS using a heavy metal ion beam, followed by gentle ablation of the imaged surface layer using C_{60}^+ ions. From the series of two-dimensional images, the three-dimensional image of triacyl glyceride distribution (m/z 815–960) was reconstructed with a volume of 1200 x 1200 x 75 μm^3. The lateral resolution in this experiment was lower than in the *Tetrahymena* example because the cells were not frozen and therefore exhibit some flexibility when bombarded. Nevertheless, it shows the applicability of mass spectrometry to whole cells in almost physiological environments.

29.1.5.8 Quantitative mass spectrometry

Mass spectrometry is not an inherently quantitative method because the ionization efficiency is different for different molecules. Consequently, the intensity of peaks in a mass spectrum is not a good measure for the abundance of the respective species. Hence, quantitative interpretation of mass spectra requires an internal reference. The term *quantitative mass spectrometry* refers to several proteomic techniques that quantify proteins from cells of different tissues or in different physiological states relative to one another or, in favorable cases, on an absolute scale.

In the simplest case we know the nature of the peptide that we wish to quantify and its sequence and mass. An isotopically labeled peptide of the same sequence, added to the sample at a known concentration, can then serve as an *internal standard*. To quantify a peptide with the natural isotopes ^{12}C and ^{14}N, a reference peptide containing the isotopes ^{13}C and ^{15}N is synthesized. The two peptides are called *isotopologs* – physico-chemically identical molecules that differ only in their isotope composition. Target and reference peptide co-elute during LC pre-fractionation of the sample, and are detected simultaneously as two peaks with different m/z by MS. Their ionization properties are also the same, and the peak intensities for the two peptides in the mass spectrum are therefore equal to their relative abundances. Since the concentration of the reference peptide is known, the absolute quantity of the target peptide in the sample can be calculated from the relative peak intensities.

In SILAC, *stable isotope labeling by amino acids in cell culture*, a cell culture to be analyzed is fed with modified amino acids, *e.g.* ^{13}C-Lys, to incorporate the isotope ^{13}C into proteins, a process termed *metabolic labeling*. A control cell culture is grown under identical conditions, but provided with the natural amino acid, *e.g.* ^{12}C-Lys. Both cells therefore produce isotopologous proteins. Equal amounts of protein extracts from labeled and non-labeled cell cultures are then mixed and analyzed by tandem LC-MS. The ratio of the peak intensities of the isotopologs equals the relative abundance of the proteins. As a control, cell cultures that are grown under identical conditions must have a ratio of unity for the isotopologs. By comparing the ratios of isotopologs between cell cultures grown under different conditions, changes in the abundance of proteins and the protein composition of the cell (the proteome) due to the different conditions can be detected. SILAC allows the effect of stress conditions, chemical treatment, genetic manipulation, or mutations in a cancer cell line to be analyzed on protein levels. SILAC depends on the isotope-labeled reference and is therefore limited to cell cultures, whose growth is time-consuming and costly. Samples taken directly from tissues cannot be analyzed by SILAC because they lack the isotope-labeled reference.

Mass spectrometry using *isobaric tags for relative and absolute quantification* (iTRAQ) introduces the internal standard by chemically modifying the sample with an iTRAQ reagent, which circumvents the requirement for metabolic labeling. The iTRAQ reagent consists of three parts: a charged reporter group, a neutral balance group, and a peptide-reactive group (Figure 29.14). Proteins are extracted from the samples that are to be compared, the total protein concentration is determined using a standard method (Section 21.2.7), and peptides are generated by proteolysis of equal amounts of each protein extract. The peptide mixtures from each sample are then reacted with an iTRAQ reagent that modifies the primary amino groups of N-termini and Lys side chains (see Section 29.3.4.1), leaving only the reporter and balance groups on the protein (Figure 29.14). A different iTRAQ reagent with a different mass distribution between reporter and balance groups is used for each peptide mixture.

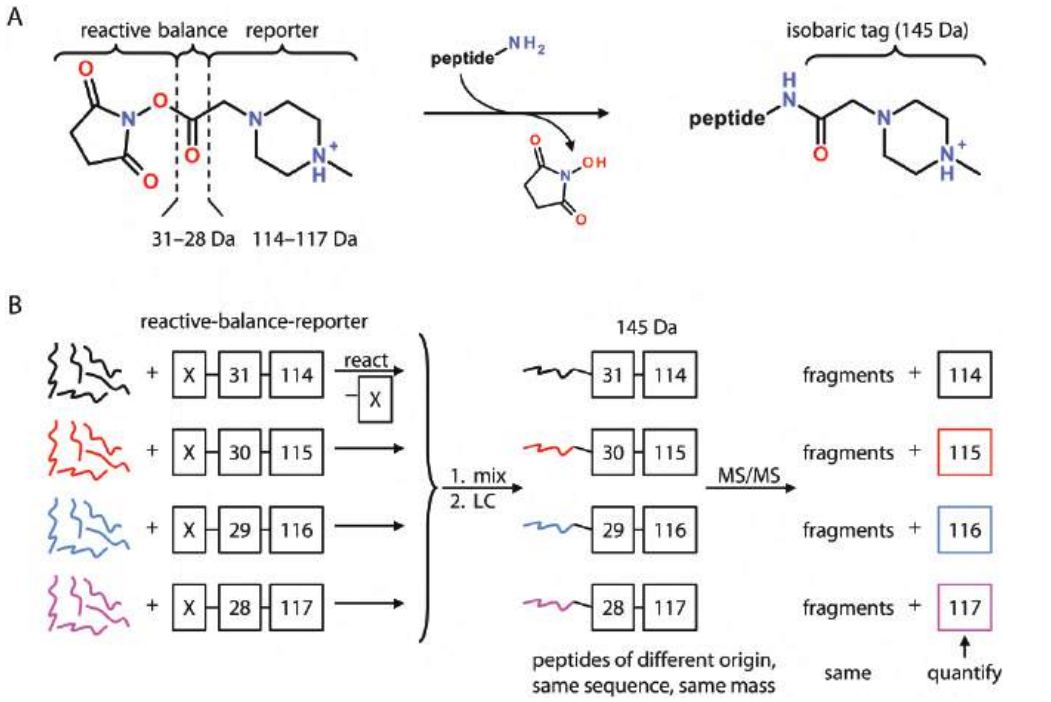


FIGURE 29.14: iTRAQ principle. A: The reagent reacts with primary amino groups *via* its activated N-hydroxysuccinimidyl ester, which transfers the reporter group, here a charged N-methylpiperazine, and the balance group, here a carbonyl group, to the peptides of the sample. By using groups with different isotope compositions, *m/z* of the reporter group is varied between 114 Da and 117 Da. The balance accordingly must range in mass between 31 Da and 28 Da, such that the total mass of the tag is constant (145 Da). B: Peptide mixtures from four sources are labeled with four different iTRAQ reagents, then mixed, and the peptides separated by LC. Peptides of the same mass but with different balance-reporter groups are analyzed by MS/MS. Their fragmentation gives the same peptide fragments, but different reporter ions, which are quantified to reflect the relative abundances of the proteins in the original cell extracts.

The iTRAQ reagents are designed such that the sum of the masses of the reporter and balance group transferred to the peptide is the same in all reagents (*isobaric tag*), but their isotope composition is different. The labeled peptides from all three samples are pooled and jointly analyzed by tandem LC-MS. Identical peptides from the different samples elute from the LC at the same time and are therefore ionized together and with the same efficiencies. Peptide sequences are determined by fragmentation to identify the protein they originated from (Section 29.1.2.4). Also, the isobaric tag decomposes in the mass spectrometer into the balance and reporter groups. The reporter groups are now not isobaric any more but have different masses for the different samples. The peak intensities of the reporter groups are measured and reflect the relative abundances of the *same* peptide from *different* samples. If the instrument is calibrated with a known concentration of the reporter groups, the peptides in all samples can be quantified on an absolute scale.

With improvements of chromatographic techniques (nano-HPLC for exact retention times), higher reproducibility from run to run in MS analysis (no drift in *m/z*), and more sophisticated data analysis software, MS moves toward a quantitative method without the requirement of internal standards, and thus to label-free quantitative MS. Label-free experiments require less amount of sample and less preparation time. The data are therefore often acquired more quickly and interpreted more easily. However, label-free MS requires more careful control of the experimental parameters than quantification based on internal standards. In label-free MS, quantification is based on chromatographic peak areas, which requires highly reproducible peptide separation between samples prior to MS analysis. By calibrating the spectrometer using increasing concentrations of a known protein, its unknown concentration in a complex sample can be quantified. This method relies on the observation that the signal intensity from ESI correlates with ion concentration. Relative abundances

of unknown proteins from different samples can be detected by comparing results from MS runs performed under identical conditions, which is not always possible. A more reliable method for label-free quantification is *spectral counting*, which relates the number of mass spectra obtained for a specific protein in many tandem MS/MS experiments to its abundance. This works because an increase of protein abundance will also increase the number of unique proteolytic peptides, which in turn leads to an increase in spectral count. Comparison of the spectral counts gives the relative protein abundances from different samples.

29.2 ANALYTICAL ULTRACENTRIFUGATION

Centrifuges have been used for hundreds of years. Here we will concentrate on the analytical uses of centrifugation by observation of the sedimentation behavior of particles, from small molecules of only 100 Da to cellular organelles reaching 10 GDa molecular mass. In analytical ultracentrifugation (AUC), a solution is subjected to a strong gravitational field. The solution components will sediment and establish a concentration gradient that is opposed by diffusion. The resulting time-dependent shape of the concentration gradient (Figure 29.15) is detected and analyzed to determine the hydrodynamic properties of a macromolecule, its molecular mass, and interactions between components of the solution. The analysis of concentration gradients is based on the forces that are experienced by the molecules, and requires no additional assumptions.

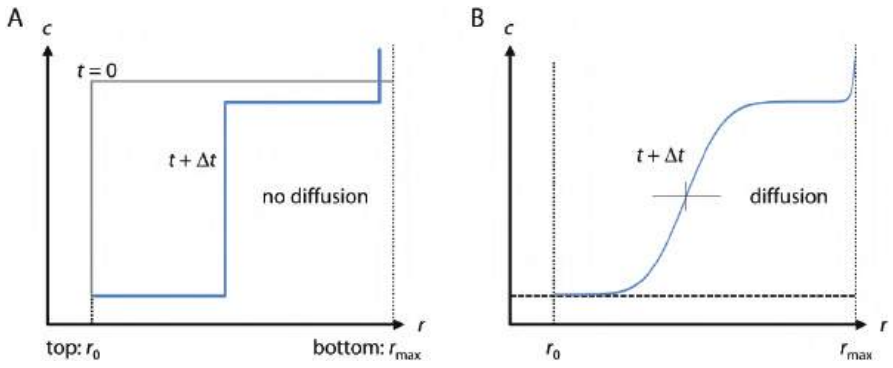


FIGURE 29.15: Change of concentration in a sample upon centrifugation. A: A sample once uniformly distributed (gray line) from the top (r_0) to the bottom (r_{max}) of the centrifuge tube is centrifuged, and the concentration is measured at a time point $t+\Delta t$ as a function of r. In the absence of diffusion, molecules sediment with a defined boundary toward the bottom of the cell, leading to a step gradient. B: Diffusion counteracts sedimentation and spreads the boundary region.

AUC experiments can be performed in two modes, termed *sedimentation velocity* and *sedimentation equilibrium*. In sedimentation velocity experiments, the molecules are sedimented at high rotor speeds, and the rate of the movement of the *boundary* is measured. From this rate, the *sedimentation coefficient* is calculated, which allows conclusions about purity, size, shape, dispersity, and molecular mass of particles. In sedimentation equilibrium experiments, rotor speeds are chosen such that sedimentation and diffusion balance each other over a time frame of hours to days. This establishes a stable concentration profile that can be analyzed with respect to molecular mass. For interacting particles, the association constant and complex stoichiometry can be determined. Before we discuss the analysis of sedimentation velocity and sedimentation equilibrium experiments (Sections 29.2.3 and 29.2.4), we will begin by introducing the instrumentation and the different options to detect the concentration gradients in AUC, followed by a general treatment of the forces that are exerted on molecules in solution in a gravitational field (Section 29.2.2).

29.2.1 INSTRUMENTATION AND DETECTION SYSTEMS

In an AUC experiment, a rotor carrying a sample operates at an angular velocity ω. For the analytical purposes in AUC, a temperature-controllable fixed-angle rotor is used (Figure 29.16). The high rotor

speeds require evacuation of the rotor chamber in order to avoid friction between the rotor and air, which would heat the sample. Most importantly, the rotor must allow passage of light through the sample for measuring the concentration of the solute as a function of distance from the center of the rotor.

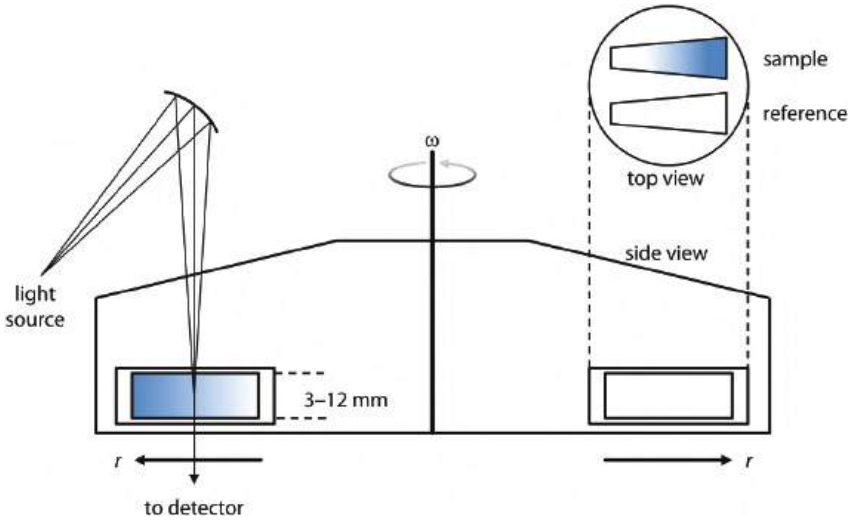


FIGURE 29.16: Schematic view of a fixed-angle AUC rotor. The angle is defined by the angle between the rotation axis and the longitudinal axis of the cell, in this case 90°. Rotors may have up to eight bores, each of which can hold a sector-shaped sample/reference cell. For absorbance measurements as a function of the cell radius *r*, light passes the sample cell from top to bottom. The standard path length of the cell is 3–12 mm.

The gravitational force generated by the rotation of the rotor leads to sedimentation of molecules in sample and reference cells (Figure 29.17) along radial paths. In parallel cells, molecules would therefore accumulate on the cell walls because of increased friction, and the resulting concentration gradient from the wall toward the center of the cell would lead to convection. The sector shape of the cell prevents accumulation of molecules on the cell walls.

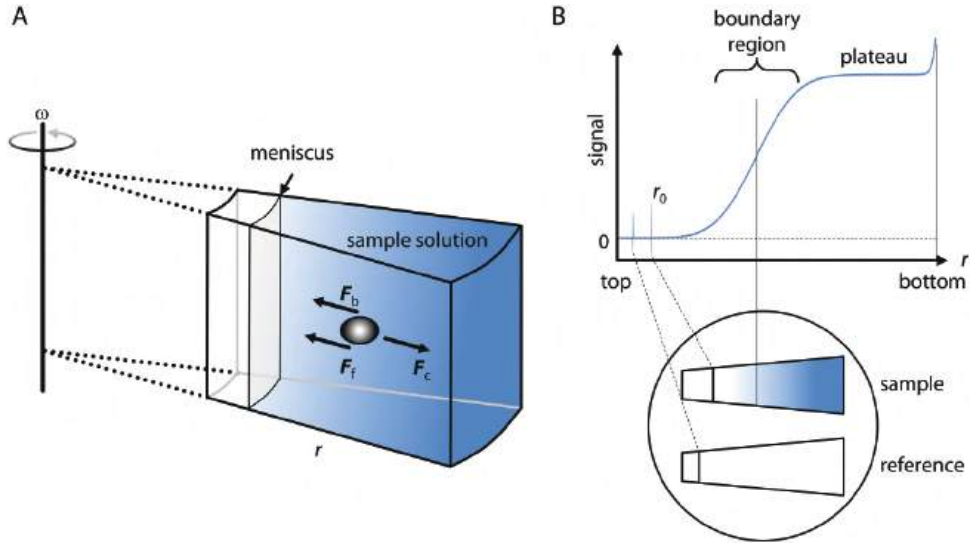


FIGURE 29.17: Principle of an AUC experiment. A: The rotor operates at an angular velocity ω, spinning a cell that holds the sample (blue). The sector shape of the cell prevents accumulation of molecules at the cell walls upon sedimentation. B: Top view of a double sector cell containing a reference (buffer) and the sample. The reference cell is filled slightly higher to distinguish one meniscus from the other (shown as disturbances in the signal above). Upon centrifugation the molecules accumulate towards the bottom of the cell, leading to a gradient that is characterized by a boundary and a plateau region.

In principle, any kind of signal from the sample analyzed is suitable to follow the sedimentation process. AUC instruments are available that measure absorption (Sections 21.2), fluorescence (Section 21.5), changes in the refractive index, or interference.

For absorption and fluorescence measurements, a suitable wavelength from a xenon lamp is selected by a monochromator. The light is directed toward the cell from the top by a grating, and transmitted light is detected by a detector below the cells. Absorption, the most frequently used analytical property in AUC, has the advantage of being directly proportional to concentration according to the Lambert-Beer law (Section 21.2.3). The presence of chromophores in proteins and nucleic acids make absorption a sensitive and selective spectroscopic probe for AUC. Given their very high extinction coefficients, nucleic acids and their complexes with proteins can be detected at micromolar concentrations.

Fluorescence typically provides higher detection sensitivity than absorbance, down to the subnanomolar range. Higher sensitivity enables the analysis of higher affinity interactions by AUC. Trp and Tyr in proteins are fluorescent, but have low quantum yields, so extrinsic fluorophores are preferred. Nucleic acids are generally non-fluorescent, and labeling with extrinsic fluorophores is required. When fluorescence is used as a probe, it must be established that the fluorescence signal is proportional to the concentration of the solute, and quenching (Section 21.5.7) or inner-filter effects (Section 21.2.8, Figure 21.56) at higher solute concentrations should be excluded.

A series of absorption or fluorescence scans of sample cells at different time points is shown in Figure 29.18.

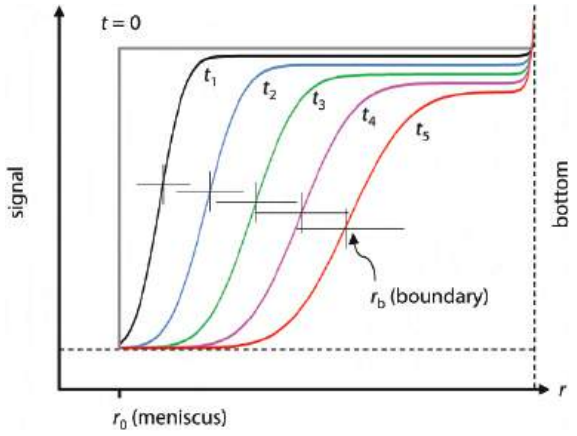


FIGURE 29.18: Spectroscopic signal as a function of radial distance is a measure of concentration. At $t = 0$ the whole cell has uniform concentration. The concentration continuously decreases close to the meniscus and increases at the bottom during centrifugation. Schematic concentration distributions are shown at constant time intervals at times t_1–t_5 after start of the centrifugation. Due to diffusion the concentration gradient covers a wider range of distances at later time points (crosses). The plateau decreases because of sedimented material. The radial position r_b is close to the inflection point of the distribution and characterizes the position of the boundary.

Solutions of polysaccharides or lipid vesicles do not contain chromophores or fluorophores for absorption or fluorescence detection. In this case, refractometric methods such as Schlieren optics or Rayleigh interference can be used. These methods measure differences in the refractive index, which are proportional to changes in concentrations.

29.2.2 Behavior of a Molecule in a Gravitational Field

The exact treatment of molecular transport processes requires thermodynamic approaches that explicitly treat Brownian motion of molecules (Section 4.1). A more intuitive understanding of the behavior of molecules in a gravitational field is obtained from a mechanical view of the forces

involved. A particle sedimenting in a gravitational field experiences three forces, the *gravitational* or *sedimenting* force F_c toward the bottom of the cell, and the *buoyancy* F_b and the *friction* F_f toward the top of the cell (Figure 29.19).

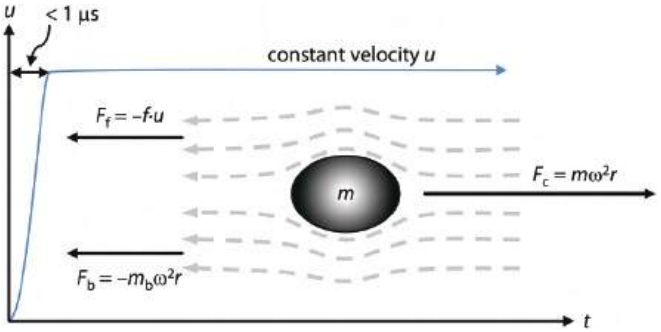


FIGURE 29.19: Forces acting on a particle in a gravitational field. A particle sedimenting in a gravitational field experiences the gravitational force F_c toward the bottom of the cell (positive sign), and the buoyancy F_b and the friction F_f toward the top of the cell (negative sign). The particle reaches a constant velocity u when these forces are in equilibrium, less than a microsecond after the start of the centrifugation. The broken arrows indicate the solvent resistance the particle experiences while moving to the right.

The gravitational (or centrifugal) force F_c generated by the centrifugal field is defined by the product of acceleration $a = \omega^2 r$ and the mass m of the particle (M_M in kg in case of a molecule).

$$F_c = m \cdot a = m \cdot \omega^2 r = m \cdot (2\pi\nu)^2 \cdot r \qquad \text{eq. 29.8}$$

The angular velocity ω is the frequency ν of rotation in revolutions per second times 2π. Division of the centrifugal acceleration by the earth gravitational acceleration $g = 9.81\ \text{m}\cdot\text{s}^{-2}$ results in the relative centrifugal force (RCF), a dimensionless number that expresses the acceleration as multiples of g.

$$\text{RCF} = \frac{\omega^2 r}{g} \qquad \text{eq. 29.9}$$

The RCF describes centrifugal forces independent of the rotational radius and hence independent of the particular rotor and centrifuge used, which facilitates comparison. For the sedimentation of biological molecules, RCFs of several hundred thousand are required.

The second force we have to take into account is the buoyancy F_b toward the top of the cell. According to Archimedes' principle, F_b is equal to the mass of solvent m_0 displaced by the submerged molecule multiplied by the gravitational acceleration, $\omega^2 r$:

$$F_b = -m_0 \cdot \omega^2 r \qquad \text{eq. 29.10}$$

The negative sign indicates that this force acts opposite to F_c. Any submerged body displaces a volume of solvent equal to its own volume. The displaced mass of solvent m_0 hence equals the product of the volume of the particle in solution and the density ρ of the solvent. The volume of the particle can be expressed as the product of the *partial specific volume* of the molecule $\bar{\nu}$ and its mass m:

$$m_0 = V \cdot \rho = m \cdot \bar{\nu} \cdot \rho \qquad \text{eq. 29.11}$$

The partial specific volume is the occupied volume of the molecule in solution per mass unit (Table 29.1, see also Section 1.7.4). Its unit is that of a reciprocal density so that the factor $\bar{\nu}\rho$ is dimensionless. As a result of the buoyancy, the mass of the molecule is reduced to an *effective mass* m_{eff} given by

$$m_{\text{eff}} = m \cdot \left(1 - \bar{\nu} \cdot \rho\right) \qquad \text{eq. 29.12}$$

From eq. 29.12, we can calculate the effective mass m_{eff} for proteins ($\bar{\nu}$ of 0.70–0.75 L kg^{-1}, Table 29.1) in aqueous solution (ρ = 1 kg L^{-1}) as 25–30% of their mass m. The effective mass of nucleic acids, in contrast, is higher, about 45% of their mass m, because of their smaller partial specific volume. We will see later that the relation in eq. 29.12 can be used to separate molecules according to their density (Section 29.2.5).

TABLE 29.1:
Partial specific volumes of macromolecules in L kg^{-1}.

Proteins	0.70–0.75
Polysaccharides	0.59–0.65
Lipids	0.95–1.0
DNA	0.55–0.59
RNA	0.47–0.55

Partial specific volumes are temperature-dependent and sensitive to solution conditions (see Section 1.7.4), and only ranges can be given.

The third and last force acting on the molecule is the frictional force F_f. Friction counteracts gravitational forces. In the absence of friction, the increase of F_c with increasing distance from the rotor axis (eq. 29.8) would continuously accelerate the molecule to higher and higher velocities. However, this is not observed because the gravitational effect is counteracted by friction (Figure 29.19), which depends on the velocity u:

$$F_f = -f \cdot u \qquad \text{eq. 29.13}$$

The *translational frictional coefficient* f depends on both the size and the shape of the molecule. Spherical molecules have the lowest frictional coefficients, whereas the friction coefficient is large for elongated, rod-shaped molecules. The frictional coefficient is a measure of how difficult it is for the solvent to flow around the molecule, or how difficult it is for the molecule to diffuse within the solvent. The Stokes-Einstein equation (eq. 29.14) relates the frictional coefficient f to the thermal energy of the molecule (Section 4.1) and the diffusion coefficient D:

$$f = \frac{k_B T}{D} = \frac{RT}{N_A D} \qquad \text{eq. 29.14}$$

(see eq. 23.13). The reciprocal relationship shows that rod-shaped molecules with large frictional coefficients diffuse more slowly than spherical molecules with lower frictional coefficients.

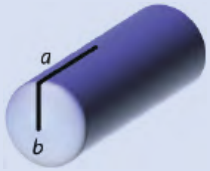
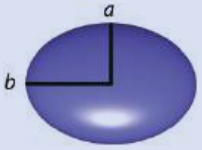
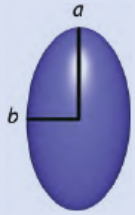
The frictional coefficient f_0 of a sphere depends on the sphere radius R_e and the solvent viscosity η. For movements of particles through continuous media (*e.g.* a protein dissolved in buffer) the relation between the frictional coefficient and the sphere radius is known as *Stokes' Law*:

$$f_0 = 6\pi\eta R_e \qquad \text{eq. 29.15}$$

(see eq. 23.14). More complicated shapes are spheroids and rods, which can serve as approximations for non-spherical proteins and elongated nucleic acids, respectively. A spheroid is an ellipse rotated about one of its axes, and can be described by the axial and equatorial half-axes a and b. Rotation of an ellipse about its major axis generates a prolate or cigar-shaped spheroid. Rotation about the minor axis leads to an oblate or lentil-shaped spheroid. A rod is a cylinder of length a and radius b. The frictional coefficients of spheroids and rods depend on their half-axes a and b (Table 29.2). It is convenient to normalize frictional coefficients with that of a sphere of equal volume (radius R_e). As

spherical molecules have the lowest friction coefficients, frictional ratios f/f_0 are always larger than unity (Figure 29.20). However, for an axial ratio < 2, oblate and prolate spheroids have similar frictional ratios that are not too different from unity. Globular proteins have frictional ratios of $f/f_0 \approx 1.2$.

TABLE 29.2:
Frictional coefficient ratios.

shape	f/f_0	R_e
long rod 	$\frac{\sqrt[3]{\frac{a^2}{b^2}}}{\sqrt[3]{\frac{3}{2}}\cdot\left(ln\left(2\cdot\frac{a^2}{b^2}\right)-0.3\right)}$	$\sqrt[3]{\frac{3a\cdot b^2}{2}}$
oblate spheroid 	$\frac{\sqrt{\frac{b^2}{a^2}-1}}{\sqrt[3]{\frac{b^2}{a^2}}\cdot tan^{-1}\left(\sqrt{\frac{b^2}{a^2}-1}\right)}$	$\sqrt[3]{a^2\cdot b}$
prolate spheroid 	$\frac{\sqrt{\frac{a^2}{b^2}-1}}{\sqrt[3]{\frac{a}{b}}\cdot ln\left(\frac{a}{b}+\sqrt{\frac{a^2}{b^2}-1}\right)}$	$\sqrt[3]{a\cdot b^2}$

For the rod, a is half of the length and b is the radius. For the spheroids, a and b are axial and equatorial half-axes, respectively. R_e is the radius of a sphere that would have the same volume as the rod or spheroid.

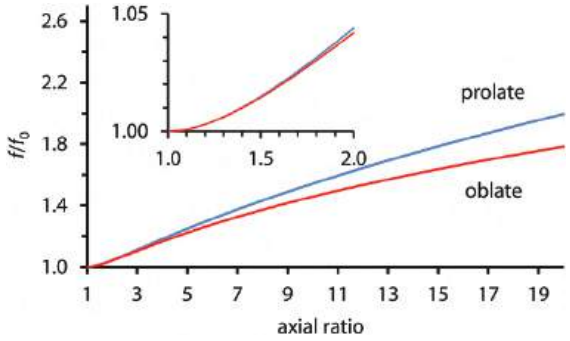


FIGURE 29.20: The frictional coefficient of spheroids depends on the shape of the particle. For particles with an axial ratio up to 2, oblate and prolate spheroids behave very similarly to each other (inset), and their frictional ratio f/f_0 is not too different from that of a sphere of equal volume (deviations < 5%).

Frictional coefficients for irregular shapes can be approximated by applying *hydrodynamic bead modeling*, which divides a trial shape into segments and assigns a (usually spherical) bead to each segment (Figure 29.21).

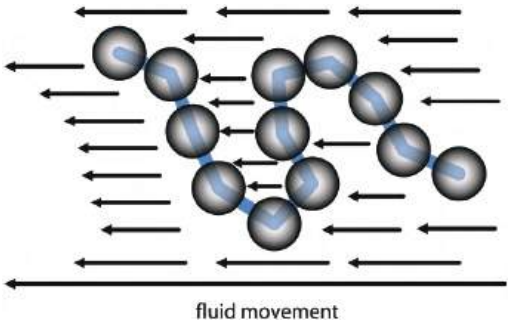


FIGURE 29.21: Calculation of frictional coefficients for irregular shapes using beads. A macromolecule is divided into segments, each of which is assigned a spherical bead with known radius. Any shape can thus be approximated as a string of beads. Some beads influence the movement of others by blocking their paths (indicated by smaller arrows).

For N segments we have N spheres, each of frictional coefficient $f_1 = 6\pi\eta R_1$. If the movement of one sphere (bead) did not influence the movement of nearby beads, the total frictional coefficient would just be the sum of the frictional coefficients of all spheres, $f_N = N{\cdot}f_1$. In reality each segment somewhat decreases the fluid movement in its vicinity. The frictional force and thus the frictional coefficient therefore become smaller than $N{\cdot}f_0$. For a molecule with N identical subunits, either a homo-oligomer or a polymer with N segments approximated by N beads, the total frictional coefficient is given by the *Kirkwood approximation* (eq. 29.16).

$$f_N = Nf_1\left(1+\frac{f_1}{6\pi\eta N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{1}{R_{ij}}\right)^{-1} \qquad \text{eq. 29.16}$$

R_{ij} is the distance between the subunit pair ij. The summation is over all pairs with $i \neq j$. The Kirkwood approximation is especially useful for symmetric homo-oligomeric complexes (Section 29.2.3). In principle, it allows the calculation of frictional coefficients for any shape by modeling the shape with a large number of spheres. However, the Kirkwood approximation is inexact for highly asymmetric particles. In such cases, it is better to determine the frictional coefficient experimentally. This can be done by measuring the sedimentation coefficient in sedimentation velocity experiments, which we will discuss in the next Section, and where the equations above are used to describe the properties of the molecules in the centrifuge.

29.2.3 Sedimentation velocity

Newton's first law of motion states that a particle rests or moves uniformly with a constant velocity in a straight direction when no external forces act upon it. We find this scenario in the ultracentrifuge after the particle has been accelerated to a constant velocity (Figure 29.19). The sum of the forces must then be zero:

$$F_c + F_b + F_f = 0 = m\cdot\omega^2 r - m\cdot\bar{\nu}\cdot\rho\cdot\omega^2 r - f\cdot u \qquad \text{eq. 29.17}$$

By expressing the mass of the molecule in eq. 29.17 as the ratio of the molecular mass M_M and Avogadro's constant N_A and rearranging we obtain

$$\frac{M_M}{N_A}\cdot\left(1-\bar{\nu}\rho\right)\cdot\omega^2 r - f\cdot u = 0 \qquad \text{eq. 29.18}$$

and

$$\frac{M_M(1-\bar{\nu}\rho)}{N_A \cdot f} = \frac{u}{\omega^2 r} \equiv s \qquad \text{eq. 29.19}$$

We define the velocity u of the particle divided by the gravitational acceleration $\omega^2 r$ as the *sedimentation coefficient* s. Sedimentation coefficients are measured in *Svedberg* (S), which is a time unit with 1 S = 10^{-13} s, or 100 fs, named after Theodor Svedberg who built the first analytical ultracentrifuge in the mid-1920s. By AUC, Svedberg clearly established that proteins were indeed macromolecules of defined molecular masses, in times where the concept of macromolecules did not yet exist! Svedberg units are frequently used in biochemistry to describe large molecular entities (Table 29.3).

According to the left-hand side of eq. 29.19, the sedimentation coefficient only depends on the properties of the particle. It is inversely proportional to the frictional coefficient and directly proportional to the *buoyant* molecular mass m_{eff} (eq. 29.12). Molecules with different shapes and sizes will therefore have different sedimentation coefficients, and there is no strict correlation between s values and molecular mass. s values are not additive because of differences in the molecular shapes of complexes compared to their constituents.

TABLE 29.3:
Sedimentation coefficients of some biological macromolecules.

particle	*s* value (S)	M_M (MDa)
prokaryotic ribosome, large subunit, small subunit	70, 50, 30	ca. 2.0 / 1.3 / 0.7
eukaryotic ribosome, large subunit, small subunit	80, 60, 40	ca. 3.2 / 2.0 / 1.2
prokaryotic ribosomal RNA (*E. coli*)	5, 16, 23	0.041 / 0.52 / 0.99
eukaryotic ribosomal RNA	18, 28	0.64 / 1.7
eukaryotic proteasome	20	0.75
spliceosome	50–60	ca. 4.8

By expressing the frictional coefficient in eq. 29.19 in terms of the diffusion coefficient (eq. 29.14), the *Svedberg equation* is obtained.

$$\frac{s}{D} = \frac{M_M \cdot (1-\bar{\nu}\rho)}{RT} \qquad \text{eq. 29.20}$$

The Svedberg equation is used for molecular mass determination after measurement of s from sedimentation velocity and D from sedimentation equilibrium centrifugation experiments, which are described in the next Sections.

29.2.3.1 Determination of sedimentation coefficients

Sedimentation coefficients can be determined from the velocity u at the known acceleration $\omega^2 r$. The velocity u is measured by following the position of the boundary as a function of time. Since the sedimentation force F_c increases with radius r, so does the velocity u. We can re-write eq. 29.19 to take the dependency of $u(r)$ into account and express u as dr/dt:

$$s = \frac{u(r)}{\omega^2 r} = \frac{\frac{dr}{dt}}{\omega^2 r} \qquad \text{eq. 29.21}$$

Separation of the variables in this differential equation prepares it for integration:

$$s\omega^2 dt = \frac{dr}{r} \qquad \text{eq. 29.22}$$

Integration from the meniscus r_0 to the radial distance of the boundary r_b yields

$$s\omega^2 t = \int_{r_0}^{r_b} \frac{dr}{r} = \ln\left(\frac{r_b}{r_0}\right) \qquad \text{eq. 29.23}$$

or

$$s\omega^2 t + ln(r_0) = \ln(r_b) \qquad \text{eq. 29.24}$$

A plot of $\ln(r_b)$ *versus* time gives a straight line with slope $s\omega^2$ from which the sedimentation coefficient of the solute can be derived.

Instead of confining data analysis to the movement of the boundary, sedimentation coefficients can also be measured by following the whole concentration distribution over time, *i.e.* the concentration of solute as a function of both time and radial distance. The diffusion coefficient D and the sedimentation coefficient s are obtained by globally analyzing the set of concentration distributions at different times with the *Lamm equation* (eq. 29.25):

$$\frac{\partial c}{\partial t} = \frac{1}{r} \cdot \frac{\partial}{\partial r}\left(r \cdot D \cdot \frac{\partial c}{\partial r} - s\omega^2 r^2 c\right) \qquad \text{eq. 29.25}$$

The Lamm equation is a differential equation that results from a combination of Fick's law of diffusion (Section 4.1) and the Svedberg equation, and provides a general description of any AUC experiment, both sedimentation velocity and sedimentation equilibrium. It takes into account *radial dilution* due to the sector-shape of the cells, as well as end effects due to accumulation of material at the bottom. If we execute the $\partial/\partial r$ differential in eq. 29.25 we arrive at a form of the Lamm equation that assumes that s and D do not depend on the concentration and that nicely shows the separate diffusion and sedimentation terms:

$$\frac{\partial c}{\partial t} = D \cdot \left(\frac{\partial c}{r \cdot \partial r} + \frac{\partial^2 c}{\partial r^2}\right) - s\omega^2 \cdot \left(2c + r \cdot \frac{\partial c}{\partial r}\right) \qquad \text{eq. 29.26}$$

The Lamm equation has no analytical solution, but various approximations to the Lamm equation have been used in AUC data analysis. Modern AUC sedimentation velocity analysis software such as Sedfit numerically solves the Lamm equation to extract s and D. To deduce the possible shape of the solute molecule, the derived values for s and D can be compared to those calculated for various trial shapes by hydrodynamic bead modeling (Figure 29.21 and eq. 29.16) using programs such as Hydropro.

29.2.3.2 Solvent and concentration dependence of the sedimentation coefficient

The value of s is not only characteristic of the molecular shape, but also depends on the viscosity and density of the solvent and the temperature at which the experiment was conducted. Measured s_{obs} values in a certain solvent and at a certain temperature are therefore often converted into a standard $s_{20,w}$ value, which corresponds to the sedimentation coefficient at 20°C in water as reference solvent according to eq. 29.27:

$$s_{20,w} = s_{obs} \cdot \frac{1 - \bar{\nu}\,\rho_{20,w}}{1 - \bar{\nu}\,\rho_{obs}} \cdot \frac{\eta_{obs}}{\eta_{20,w}} \qquad \text{eq. 29.27}$$

$\rho_{20,w}$ is the density of water at 20°C and ρ_{obs} that of the solvent at the temperature of the measurement. Likewise, $\eta_{20,w}$ is the viscosity of water at 20°C and η_{obs} the viscosity of the solvent at the temperature of the measurement.

The value of *s* also depends on the solute concentration, and *s* becomes smaller with higher concentration. One reason for this decrease in *s* is the slowdown of sedimentation due to the increased viscosity of the concentrated solution. In addition, sedimenting particles must continuously displace solvent during sedimentation. At high concentrations, they also have to displace other solute macromolecules (molecular crowding), which are more difficult to displace than the solvent. This effect is larger for extended molecules with high frictional coefficients such as nucleic acids, unfolded proteins, or glycosylated proteins. The concentration dependence of the sedimentation coefficient follows a hyperbolic relation (eq. 29.28):

$$s_{20,w}(c) = \frac{s_{20,w}(c=0)}{(1+k_s \cdot c)} \qquad \text{eq. 29.28}$$

The constant k_s has units of inverse concentration and depends on the shape of the molecule. It is usually very small for globular proteins (ca. 0.005 mL·mg^{-1}) but larger for extended molecules (Figure 29.22). Extrapolation of $s_{20,w}$ to zero concentration is useful for comparing *s* values from experiments performed at different concentrations.

The concentration dependence of *s* may lead to sharpening of the boundary region. Molecules preceding the boundary r_b are slowed down because of the higher concentrations whereas molecules in the trailing edge of the boundary can move faster. The net effect is a partial counteraction of diffusion, leading to a sharper boundary region.

A change in sedimentation coefficient with solute concentration may also reflect self-association. Such a concentration dependence of *s* values reflects the higher molecular mass and the changed shape of the complex (Figure 29.22), and can be used to determine the equilibrium association constant of the complex (Section 2.2).

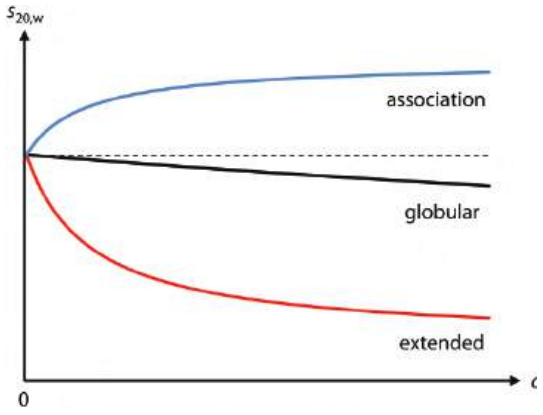


FIGURE 29.22: Concentration dependence of *s*. The $s_{20,w}$ value at zero concentration ($c = 0$) changes in a different manner with increasing concentration depending on the shape and association state of the molecule.

29.2.3.3 Measuring polydispersity and association

If there are several components in a sample, the concentration gradients measured in sedimentation velocity experiments are a superposition of the individual traces for each component. The presence of more than one component might not be apparent from the shape of the traces when the molecular masses are similar or the concentration of a component is very low. Two solutes with sufficiently different molecular masses that are both present in detectable amounts will give rise to double-sigmoidal traces (Figure 29.23). In this case, the polydispersity, *i.e.* the number of non-interacting species and their abundance, can be assessed.

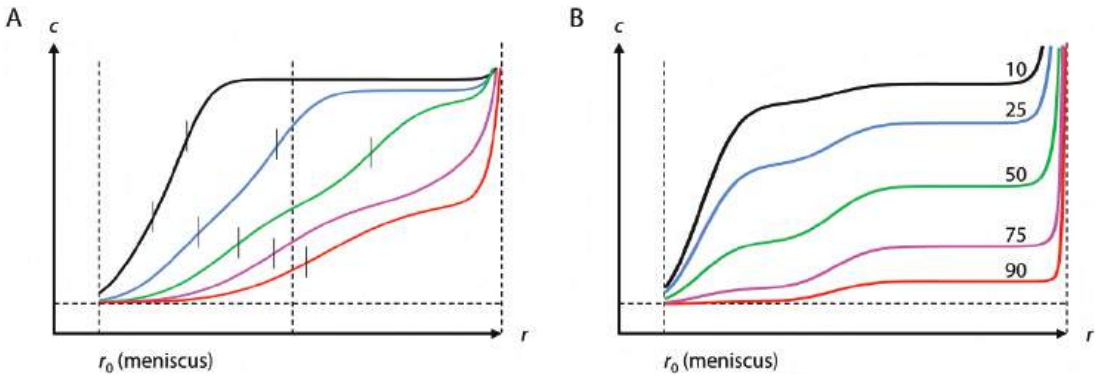


FIGURE 29.23: Simulated sedimentation velocity experiments for mixtures of two species. A: Equimolar mixture of two non-interacting molecules with molecular masses of 20 kDa and 80 kDa. Traces are calculated for different time points 5 min apart. B: Traces at one particular time point for different ratios of the two molecules. The numbers indicate the percentage of the 80 kDa species.

AUC data for mixtures can be evaluated using a linear combination of Lamm equations, one for each component, weighted by the component concentration. The analysis provides a distribution of *s* values from which the relative abundance and the *s* value of each species can be extracted (Figure 29.24).

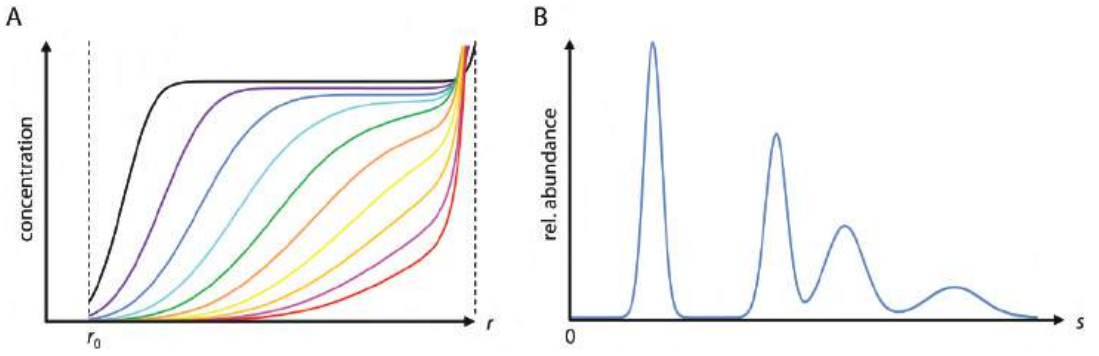


FIGURE 29.24: Sedimentation of a polydisperse mixture. A: Sedimentation data of a mixture of four components with *s* values 3 S, 7.5 S, 10 S, and 14 S. B: Distribution of *s* values obtained from analysis of the sedimentation data with a concentration-weighted sum of Lamm equations. The width of the distributions increases with increasing *s* values.

If the molecules in the solution interact, sedimentation velocity experiments provide information on complex formation. Depending on the time scale of association and dissociation reactions with respect to the time scale of the AUC experiment, each trace may contain information about any species present in the equilibrium at any time point and at any radial position. The traces obtained are therefore more complex than those from mixtures of non-interacting molecules. The simplest case is the monomer-dimer equilibrium:

$$A + A \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} A_2 \qquad \text{scheme 29.1}$$

If binding and dissociation is slow compared to the time scale of the AUC experiment, the mixture will behave like a polydisperse solution with two boundaries per trace. In this case, analysis of the traces with the Lamm equation yields the distribution of *s*, the *s* values for monomer and dimer, and their relative abundance in equilibrium. Otherwise, the traces show a single asymmetric

boundary, and analysis of the velocity of its movement gives a weight-averaged sedimentation coefficient. By measuring this weight-averaged coefficient as a function of monomer concentration, the equilibrium association constant can be determined.

29.2.4 Sedimentation equilibrium

In AUC experiments with lower rotor speeds, on the order of 10000 *g* depending on the size of the molecules, diffusion plays a major role in shaping the concentration profile of the sedimenting molecules. The rate of diffusion is proportional to the concentration gradient (Section 4.1, eq. 4.1). Therefore, diffusion is higher near the bottom of the cell where the concentration is higher, and net diffusion is thus directed toward the top of the cell. Because of the radial dependence of the centrifugation force, sedimentation is faster near the bottom of the cell, and it is directed toward the bottom of the cell (Figure 29.25). After many hours, these opposing effects establish a stable concentration gradient following an exponential function, where at each point of the profile sedimentation and diffusion are balanced.

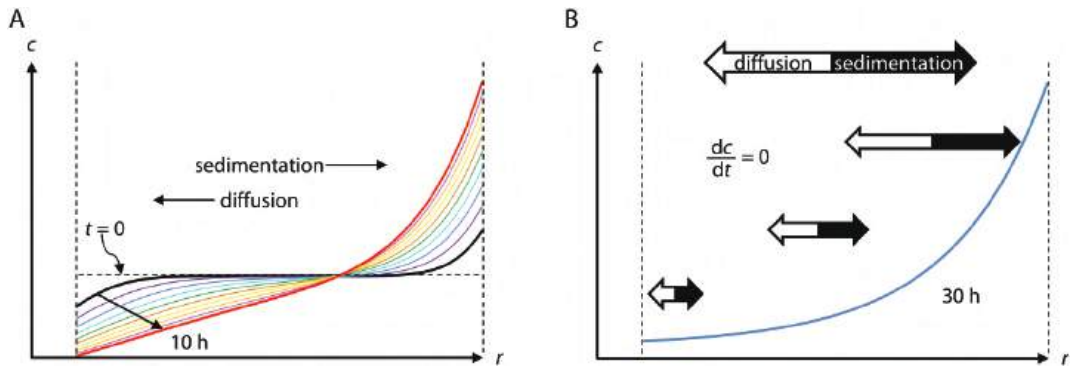


FIGURE 29.25: Approach to sedimentation equilibrium. A: Simulated traces for a 50 kDa protein centrifuged at 10000 *g*. At the outset of the experiment, the concentration of solute is constant throughout the cell (dashed horizontal line). After 1 h (black), the meniscus region has lower concentration than the bottom of the cell. The sedimentation force increases with radial distance, leading to a higher concentration at larger *r*. Diffusion opposes sedimentation because it redistributes solute to regions of smaller concentration (*i.e.* smaller *r*), leading to establishment of a stable concentration gradient over the experiment. B: Invariant solute distribution is reached after 30 h of centrifugation. The solute profile follows an exponential function. The size of the arrows shows that the magnitude of the opposing forces increase at higher concentrations.

Both sedimentation and diffusion depend on the diffusion coefficient D or the frictional coefficient f (eq. 29.14). When diffusion and sedimentation have reached equilibrium, the data obtained are independent of D and f, and thus are independent of the shape of the molecule.

29.2.4.1 Determination of molecular mass using sedimentation equilibrium

Analysis of the equilibrium concentration distribution as a function of radial position directly yields the molecular mass, which can be compared to the results from mass spectrometry (Section 29.1) or SAXS (Section 23.2.4). To derive an expression that allows determination of the molecular mass from the radial concentration dependence $c(r)$ in equilibrium, we begin by recalling the definition of the chemical potential (eq. 1.162):

$$\mu = \mu^0 + RT \ln \frac{c}{c^0} = \mu^0 + RT \ln c(r) \qquad \text{eq. 1.162}$$

c^0 is the concentration at standard conditions (1 M). μ^0 is the chemical potential at standard conditions. The solute concentration c is a function of the radial position r. The total potential μ_{tot} of the

solution in the centrifuge is the sum of the chemical potential μ and a centrifugal potential $U(r)$ that is also a function of the radial position:

$$\mu_{tot} = \mu + U(r) = \mu^0 + RT \ln c(r) + U(r) \qquad \text{eq. 29.29}$$

At equilibrium, the potential μ_{tot} is at a minimum, hence its derivative is zero. We can calculate the derivative of the sum as the sum of the individual derivatives, and obtain

$$\frac{d\mu_{tot}}{dr} = \frac{d\mu^0}{dr} + RT\frac{d\ln c(r)}{dr} + \frac{dU(r)}{dr} = 0 \qquad \text{eq. 29.30}$$

Since μ^0 is a constant, its derivative is zero and we have the following condition for equilibrium:

$$RT\frac{d\ln c(r)}{dr} = -\frac{dU(r)}{dr} \qquad \text{eq. 29.31}$$

The term $-dU/dr$ is the negative *gradient* of the centrifugal potential, which is equivalent to the centrifugal force F_c that we have defined in the preceding Section as $F_c = m\omega^2 r$ (eq. 29.8). In general, a force is the negative gradient of a potential function. Substituting the effective mass (eq. 29.12) into this equation, we obtain:

$$\frac{dU(r)}{dr} = -F_c = -m\cdot\omega^2 r = -m\cdot(1-\bar{\nu}\rho)\cdot\omega^2 r \qquad \text{eq. 29.32}$$

The chemical potential in eq. 1.162 uses molar concentrations, and in order to reflect molar energy changes of the centrifugal field, we have to multiply F_c with N_A, which gives

$$-\frac{dU(r)}{dr} = F_{c,molar} = M_M(1-\bar{\nu}\rho)\cdot\omega^2 r \qquad \text{eq. 29.33}$$

We can then combine eq. 29.31 and eq. 29.33 to

$$RT\frac{d\ln c(r)}{dr} = M_M(1-\bar{\nu}\rho)\cdot\omega^2 r \qquad \text{eq. 29.34}$$

separate the variables

$$RT\,d\ln c(r) = M_M(1-\bar{\nu}\rho)\cdot\omega^2 r\cdot dr \qquad \text{eq. 29.35}$$

and integrate from a reference concentration c_0 at radial position r_0 to any concentration c at its respective radial position r:

$$\int_{c_0}^{c} RT\,d\ln c(r) = \int_{r_0}^{r} M_M(1-\bar{\nu}\rho)\cdot\omega^2 r\,dr \qquad \text{eq. 29.36}$$

Evaluation of the integral gives

$$RT\left[\ln c(r)\right]_{c_0}^{c} = M_M(1-\bar{\nu}\rho)\cdot\omega^2\cdot\left[\frac{r^2}{2}\right]_{r_0}^{r} \qquad \text{eq. 29.37}$$

and

$$\ln\frac{c(r)}{c_0} = \frac{M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT} \qquad \text{eq. 29.38}$$

According to eq. 29.38, a plot of $\ln(c/c_0)$ as a function of r^2 is linear, and from its slope, the molecular mass M_M can be calculated. To determine the molar mass M from such a plot, the reference concentration c_0, for instance at the meniscus of the cell r_0, has to be known. With modern fitting programs, linearization is not necessary but the data can be described directly by the exponential function we obtain by solving eq. 29.38 for $c(r)$:

$$c(r) = c_0 \cdot e^{\frac{M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} \qquad \text{eq. 29.39}$$

In cases where several components of different molecular masses are present in the sample, the observed concentration distribution will be a linear combination of several instances of eq. 29.39. The dependence of $\ln(c/c_0)$ on r^2 (eq. 29.38) will then not be linear anymore, which serves as a diagnostic for polydispersity.

29.2.4.2 Association in sedimentation equilibrium

In a mixture of interacting components an equilibrium between all possible species is established at every point in the centrifuge. For the simplest case, a monomer-dimer equilibrium (Scheme 29.1), the equilibrium association constant is

$$K = \frac{c_2}{{c_1}^2} \qquad \text{eq. 29.40}$$

c_2 and c_1 are the concentrations of the dimer and monomer, respectively. Eq. 29.40 must be fulfilled at each and every point within the sample cell. We can formulate the radial dependence of the concentration (eq. 29.39) for dimer and monomer, and include the fact that the mass of the dimer is twice that of the monomer. Division of the concentration profiles gives:

$$\frac{c_2(r)}{c_1(r)} = \frac{c_2(r_0)e^{\frac{2M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}}}{c_1(r_0)e^{\frac{M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}}} = \frac{c_2(r_0)}{c_1(r_0)}e^{\frac{M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} \qquad \text{eq. 29.41}$$

The exponential term on the right-hand side is just $c_1(r)$ divided by $c_1(r_0)$ (eq. 29.39). Substituting the exponential term in eq. 29.41 with $c_1(r)/c_1(r_0)$ yields

$$\frac{c_2(r)}{c_1(r)} = \frac{c_2(r_0)}{c_1(r_0)} \cdot \frac{c_1(r)}{c_1(r_0)} \qquad \text{eq. 29.42}$$

or

$$\frac{c_2(r)}{{c_1}^2(r)} = \frac{c_2(r_0)}{{c_1}^2(r_0)} = K \qquad \text{eq. 29.43}$$

This result tells us that the equilibrium constant is the same at every position of the cell, as it should be. Furthermore, eq. 29.43 is also independent of the rotor speed ω, and the equilibrium distribution of c_1 and c_2 will be the same at any position at any rotor speed.

In the centrifuge, monomer and dimer are not observed independently. Only the total signal corresponding to a weighted average of the contributing species is observed (Figure 29.26).

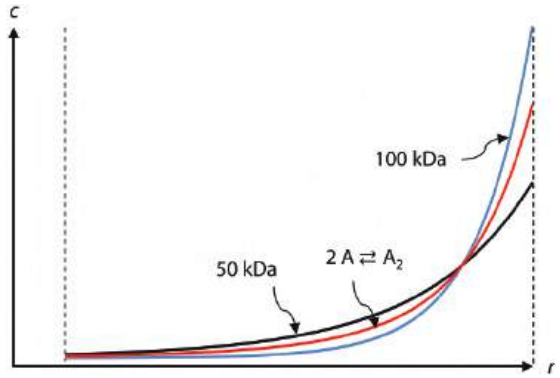


FIGURE 29.26: Concentration profile for the sedimentation equilibrium of a monomer that self-associates to a dimer. The black and blue traces are the expected sedimentation profiles of a 50 kDa monomer and a 100 kDa dimer, respectively. The red trace is the observed concentration profile for a monomer-dimer equilibrium. The equilibrium association constant of the complex can be derived by fitting a combination of the expected traces for the pure monomer and dimer to the observed data.

The dimer contributes twice to the detected trace compared to the monomer. The total absorption signal is then a weighted sum of monomer and dimer contributions according to the Lambert-Beer law:

$$A_{\text{tot}}(r) = \varepsilon_1 c_1(r) d + \varepsilon_2 c_2(r) d = \varepsilon_1 c_1(r) d + 2\varepsilon_1 \cdot K c_1^{\,2}(r) d \qquad \text{eq. 29.44}$$

Here, d is the path length of the cell and $\varepsilon_1 = 0.5\varepsilon_2$ represents the assumption that the extinction coefficient of the dimer is twice that of the monomer. The right-hand side of eq. 29.44 just expresses c_2 by c_1 using eq. 29.40. By expressing $c_1(r)$ and $c_2(r)$ according to eq. 29.39, the total signal $A_{\text{tot}}(r)$ can be written as

$$A_{\text{tot}}(r) = \varepsilon_1 c_{1,r_0} d \cdot e^{\frac{M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} + 2\varepsilon_1 K c_{1,r_0}^2 d \cdot e^{\frac{2M_M(1-\bar{\nu}\rho)\omega^2\left(r^2-r_0^2\right)}{2RT}} \qquad \text{eq. 29.45}$$

Using eq. 29.45, we can calculate the equilibrium association constant from the measured absorption profile, provided that the concentration of the monomer at a reference point r_0 and the molecular mass of the monomer are known. We see that the derived expression is rather complex even for the comparatively simple monomer-dimer equilibrium. Equations for equilibria between more than two components and for complex stoichiometries different from unity have been derived and implemented in AUC fitting programs such as Sedfit. Sedimentation equilibrium experiments are suitable for measuring a broad range of equilibrium association constants (10^4–10^8 M^{-1}).

29.2.5 Zonal, band, or isopycnic centrifugation

Concentration gradients such as those established in sedimentation equilibrium experiments can also be used for preparative purposes to separate molecules according to their density. When the

densities of the solute ($1/\bar{v}$) and the solvent (ρ) are equal (*isopycnic*), the factor $\bar{v}\rho$ in eq. 29.12 is unity and the effective mass m_{eff} of the particle is zero. In this situation, no net force acts on the molecule, there is no tendency to move, and the sample is concentrated in a *zone* or *band* at the corresponding position of the density gradient.

Density gradients can be prepared in continuous and discontinuous forms. For *step gradients*, a highly concentrated solution of sucrose or the polysaccharide ficoll is overlaid with layers of less and less concentrated solutions (Figure 29.27). In contrast, gradual mixing of glucose, sucrose, or glycerol solutions of different density delivers a *continuous gradient* of desired steepness (Figure 29.27). Such gradients are then used to fractionate by centrifugation components in complex mixtures according to their density. Both types of gradient are stable for a number of hours. The centrifugation is usually performed in swing-out rotors to minimize convection. In step gradients, the particles accumulate on top of the first layer with a higher density than the density of the particle. In continuous gradients the sedimenting particles form bands at positions where their density equals that of the gradient. The fractions are recovered by piercing the tubes with a syringe at the position of the band, or by drop-wise collection (Figure 29.27). Gradient centrifugation allows fractionation of different blood cell types from blood plasma, and cellular organelles or large macromolecules such as ribosomes and ribosomal subunits from cell extracts.

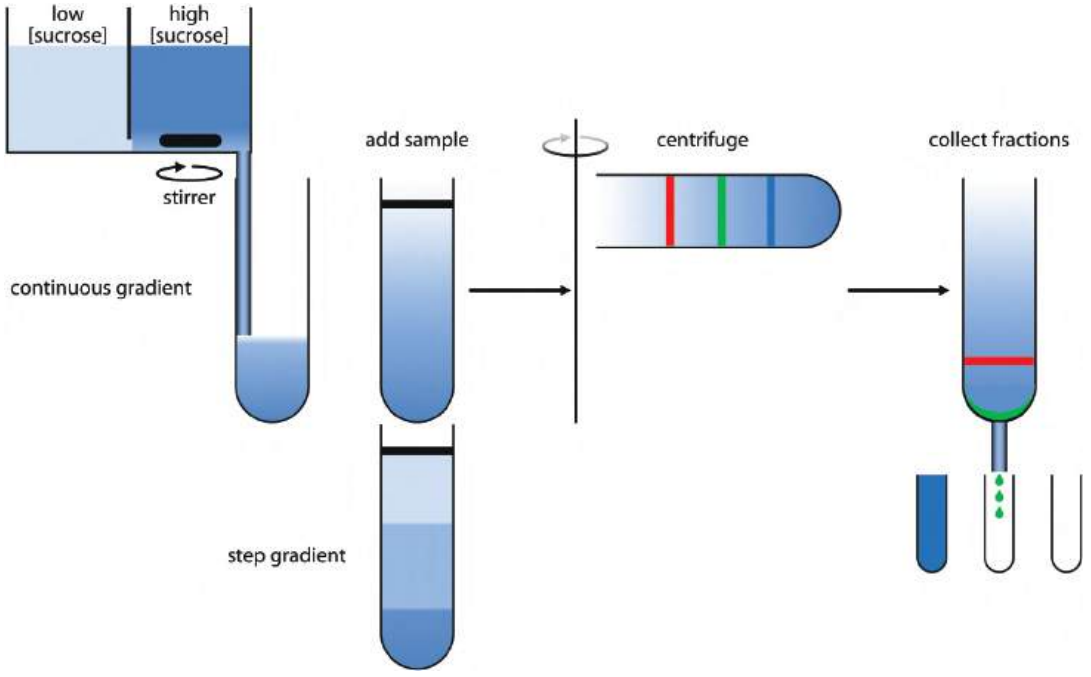


FIGURE 29.27: Principle of sucrose gradient centrifugation. A continuous sucrose gradient is prepared in a gradient mixer by mixing of two sucrose solutions of different concentrations. Alternatively, a step gradient can be generated by overlaying sucrose solutions of different densities layer by layer. The sample, of lower density than the top of the gradient, is applied to the gradient surface. The gradient is centrifuged in a swing-out rotor to separate the components of the sample into distinct bands. Fractions are recovered by piercing the tube to extract the band of interest, or by collecting fractions from bottom to top.

Instead of manual mixing of gradients, sedimentation equilibrium can be used to generate concentration gradients *in situ* during the separation (see Figure 29.25). In this type of experiment, the sample is mixed with a salt solution, typically CsCl or Cs_2SO_4, and centrifuged for several hours. During the centrifugation, the density gradient is established, and the particles move to their

isopycnic positions. The concentration distribution of a pure solute centrifuged in a density gradient follows a symmetric Gaussian function of the radial position (Figure 29.28):

$$c_{\text{solute}}(r) = c_{\text{solute}}(r_{\text{eq}})\, e^{-\frac{(r-r_{\text{eq}})^2}{2\sigma^2}} \qquad \text{eq. 29.46}$$

The gradient has a concentrating effect because of the balancing of diffusion and sedimentation forces, which have different signs on either side of the equilibrium position r_{eq} (Figure 29.28). The standard deviation σ of the Gaussian (at 68% of the peak area) is inversely related to the molecular mass of the solute:

$$\sigma^2 = \frac{RT}{M_{\text{M}}\bar{\nu}\left(\frac{\text{d}\rho}{\text{d}r}\right)_{r_{\text{eq}}}\omega^2 r_{\text{eq}}} \qquad \text{eq. 29.47}$$

Thus, measuring the width σ and position r_{eq} of the band allows calculation of the molar mass of the solute, provided that the steepness of the gradient is known.

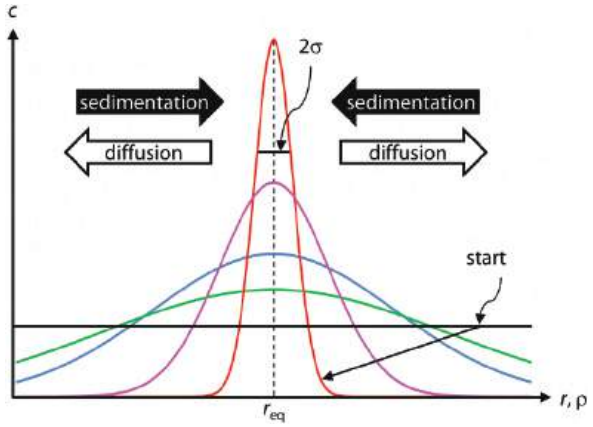


FIGURE 29.28: Band formation in linear gradients. At the outset, the cell contains a homogeneous mixture of solute (black line). During centrifugation the solute molecules are concentrated at a radial position r_{eq} that equals their buoyant density (dashed line). The opposite directions of movement due to sedimentation and diffusion lead to a sharpening effect over the time of the experiment (green, blue, magenta, and finally red curve).

Isopycnic centrifugation in self-establishing gradients is particularly successful for the analysis of nucleic acids that form highly focused bands because of their high frictional coefficients and hence very slow diffusion. CsCl density gradients allow the separation of AT-rich and GC-rich DNA, or of supercoiled plasmids from contaminating RNA. Historically the most prominent contribution of CsCl gradient centrifugation was the proof that DNA replication is semi-conservative (Box 29.1).

BOX 29.1: SEMI-CONSERVATIVE DNA REPLICATION – THE MESELSON-STAHL EXPERIMENT.

When in 1953 James Watson and Francis Crick, based on the X-ray diffraction data of Rosalind Franklin, published the structure of the DNA double helix, a replication model for DNA was immediately apparent. Each single strand of DNA can serve as a template to reproduce the other strand. This model predicts that the two new DNA double helices

(*Continued*)

each contain one "old" and one "new" strand. Using CsCl gradient centrifugation, Matthew Meselson and Frank Stahl proved this in 1958 (Figure 29.29). *E coli* was grown on medium with $^{15}NH_4Cl$ as the only nitrogen source for several generations to produce a heavy version of its DNA. The bacteria were then quickly shifted to a medium containing the usual ^{14}N isotope. DNA samples were analyzed by centrifugation in a time-dependent manner. After one generation of bacterial growth (ca. 20 min), only a single band with a density exactly between that of the $^{15}N/^{15}N$ (both strands ^{15}N-labeled) and $^{14}N/^{14}N$ double-stranded DNA was detected, which corresponded to an $^{14}N/^{15}N$ hybrid DNA. No parent DNA ($^{15}N/^{15}N$) was detected anymore. This proved that all daughter DNA inherited exactly half of the parent DNA. After two generations, two bands of equal concentration were observed, corresponding to $^{14}N/^{15}N$ and $^{14}N/^{14}N$, which is also in agreement with semi-conservative replication.

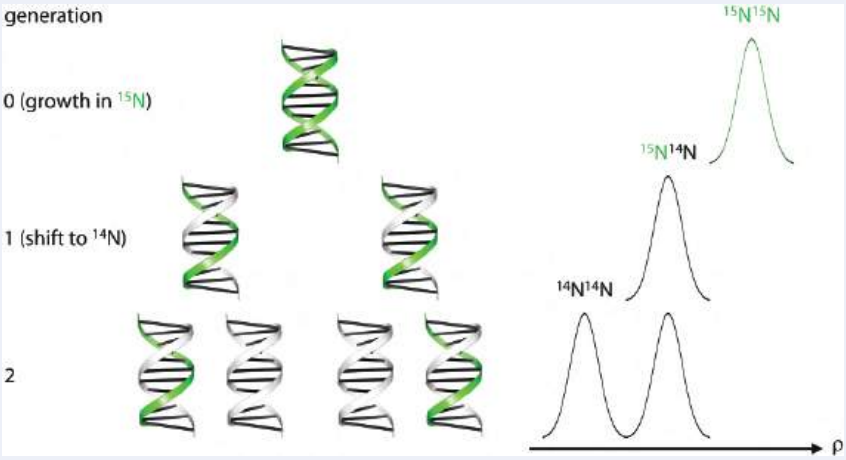


FIGURE 29.29: The Meselson-Stahl experiment. Bacteria are grown in medium with $^{15}NH_4Cl$ to label both strands of their DNA with ^{15}N (green; generation 0). In generation 1, grown on medium with ^{14}N, the $^{15}N/^{15}N$ DNA is converted into a DNA with a density between $^{15}N/^{15}N$ and $^{14}N/^{14}N$ (detected by density gradient centrifugation, right), consistent with the formation of a $^{14}N/^{15}N$ hybrid (green/gray). In generation 2, equal amounts of $^{14}N/^{15}N$ and $^{14}N/^{14}N$ DNA are observed, in agreement with predictions from the model of semi-conservative DNA replication.

The density of proteins in CsCl solutions is observed to change with pH. The increased number of negative charges at higher pH leads to binding of Cs^+ ions near carboxylates. Deprotonation of His and Lys, on the other hand, will release less dense, hydrated anions. Both effects increase the density of the protein. The extent of salt binding to proteins may also change depending on their final position in the gradient. In general, due to interaction with the gradient material, which is a situation far away from an ideal solution in a thermodynamic sense, no precise molecular masses can be determined. A very useful application for this equilibrium density gradient centrifugation is the detection of *changes* in densities due to protein-nucleic acid or protein-lipid complex formation. The partial specific volumes of these components are quite different (Table 29.1), leading to measurable changes in density although the molecular masses may not change dramatically.

29.3 SURFACE PLASMON RESONANCE

Surface plasmon resonance (SPR) has widespread applications in the biophysical characterization of molecular interactions. The principle of the method is rather simple. A *receptor* molecule is fixed to a sensor surface and brought into contact with a binding partner, termed the *analyte*. The refractive index near the surface changes upon binding, and this change is measured in a time-dependent fashion. The SPR signal contains kinetic and thermodynamic information: association and

dissociation rates, affinities, and stoichiometries of binding can be obtained. SPR measurements are fast and can be automated, which makes them particularly useful for screening of small molecules for binding to pharmaceutical drug targets.

29.3.1 Physical background of SPR

The SPR phenomenon can be understood starting from total internal reflection (TIR) at an interface of two non-absorbing materials (see Figure 25.10 in Section 25.4). Light propagating from a medium of high refractive index n_1 to a medium of low refractive index n_2 will be either refracted or reflected, depending on the incidence angle θ relative to the optical axis (Figure 29.30). At small angles θ the light is refracted away from the optical axis. Once the incidence angle θ matches the *critical angle* θ_c, the light will be refracted perpendicular to the optical axis and then propagates parallel to the interface. At higher incident angles θ, total internal reflection occurs. The fully reflected beam leaks electric field intensity, called an *evanescent wave*, into the medium of lower refractive index. The evanescent wave amplitude decreases exponentially with increasing distance from the interface (see eq. 25.20, Section 25.4).

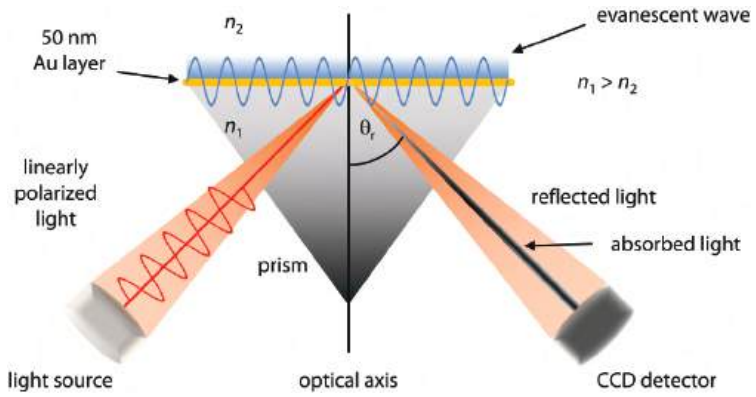


FIGURE 29.30: Principle of surface plasmon generation. Linearly polarized light is coupled into a glass prism, such that it impinges on the glass-sample interface at different angles θ with respect to the optical axis. At a certain angle θ, the light is totally reflected at the interface. If a gold layer is placed at the n_1/n_2 interface, the evanescent wave excites a surface plasmon (blue sine wave) in this layer, which leads to a slight decrease of the light intensity in the reflected beam.

In SPR, the evanescent wave generated by total internal reflection is allowed to interact with an electrically conducting material, a thin metal layer that is placed at the glass-sample interface. Gold is frequently used because it is chemically almost inert, but silver and aluminium also work well. The incident light has vector components perpendicular and parallel to the plane of the metal layer, and causes displacement of the metal electrons in both directions. The lateral movement of electrons due to displacement in the plane of the metal layer is observed as electrical conductivity. In contrast, displacement perpendicular to the plane of the metal layer increases the distance between electrons and nuclei (that are fixed in the plane), leading to charge separation. The vector component of the electromagnetic field perpendicular to the metal layer oscillates, and the electrons therefore also oscillate, giving rise to a *charge density wave* of the same wavelength as the incident light (Figure 29.30). The energy quantum of this charge density wave is termed a surface *plasmon*, in analogy to the energy quantum *photon* for a free electromagnetic wave. The excitation of the surface plasmon leads to a reduction of the light intensity in the reflected beam, which is measured in an SPR experiment using a CCD detector (Figure 29.30). The energy of the surface plasmon itself is eventually dissipated into heat.

The intensity of the surface plasmon at the metal-glass prism interface in an SPR experiment depends on the magnitude of the vector component of the incident light that is perpendicular to the metal layer. This component is maximized by using laser light polarized in the *plane of incidence*,

which is the plane that is defined by the optical axis and the propagation direction of the laser beam. In addition, the angle of incidence θ needs to be optimized by rotating the prism with respect to the laser beam (Figure 29.31). The *resonance angle* θ_r with maximal intensity of the surface plasmon depends on the laser wavelength and the refractive indices n_1 and n_2. At this angle, the minimal light intensity of the reflected beam is detected.

The amplitude of the surface plasmon decreases exponentially with distance from the glass-sample interface. The plasmon extends only about 1 λ (a few hundred nm) away from the gold layer into the medium (the sample) of refractive index n_2. Within the narrow layer near the surface that is probed by the surface plasmon, SPR is therefore sensitive to changes in the refractive index n_2 of the sample. This property makes surface plasmons useful as optical probes to follow binding events on surfaces. Binding of a molecule to a partner on the surface leads to an increase in mass and to a change in the refractive index n_2 near the surface. This change in refractive index in turn causes a change in the resonance angle θ_r at which the surface plasmon intensity is maximal and the intensity of the reflected beam has a minimum (Figure 29.31). During an SPR experiment, the change in resonance angle is measured as a *response unit* (*RU*). SPR experiments can detect changes in SPR angle by 0.001°, which corresponds to changes in n_2 by as little as 10^{-5}. Because of the local confinement of the surface plasmon, SPR is insensitive to properties farther away from the surface, so that samples may be colored, opaque, or turbid.

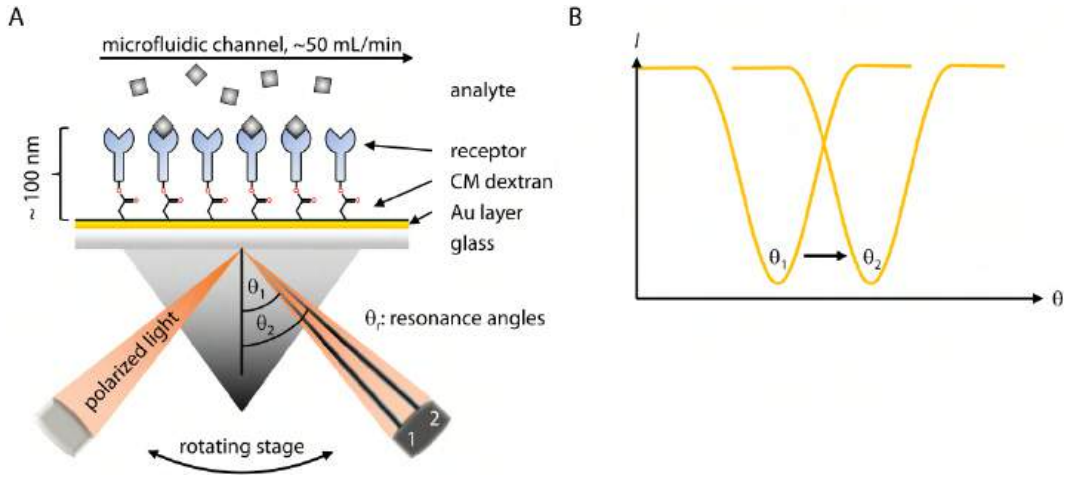


FIGURE 29.31: Principle of an SPR experiment. A: The gold layer is functionalized with the first binding partner (receptor). A solution containing the second binding partner (analyte) is passed over the sensor chip. Binding causes a change in mass near the gold layer and changes the refractive index, which in turn alters the resonance condition θ for the surface plasmon. B: Intensity of the reflected light as a function of θ. θ_1 is the resonance angle in the absence of the binding partner, θ_2 is the resonance angle after binding. The change can either be monitored by changing the angle of incidence or by moving the prism on a rotating stage.

29.3.2 Principle and information content of an SPR experiment

In an SPR experiment, the first binding partner (receptor), typically the macromolecule, is covalently or non-covalently immobilized on the metal layer of the sensor surface of the SPR chip (Section 29.3.4). The receptor can be a small molecule, lipid, carbohydrate, protein, nucleic acid, or a large particle such as a virus or cell. The receptor is also often called "ligand" in SPR literature, which may be confusing to biochemists because ligand usually stands for a small molecule or a mobile binding partner. In the following, we will use the term *receptor* for the molecule that is attached to the sensor chip, and *analyte* for the second binding partner in solution.

The SPR measurement is carried out in a microfluidic device that allows rapid switching between solutions that are in contact with the sensor and carrying the immobilized receptor (Figure 29.32). First, a baseline of the signal (the *response*) is established by washing the sensor surface with buffer. In a *contact* or *association phase*, the sensor surface is then brought into contact with the solution

that contains the analyte, and binding is followed as an increase in the SPR signal as a function of time. Finally, the *dissociation phase* is started by flowing buffer over the sensor chip again. During the dissociation phase, bound analyte dissociates, which is recorded as a reduction in SPR signal (Figure 29.32). After dissociation has completed, the sensor chip is in principle ready for another binding reaction. Regeneration of the chip by an extra washing step is usually required to remove unspecifically bound analyte. The SPR response as a function of time is termed the *sensorgram*. The sensorgram is corrected for signals unrelated to the actual binding event, such as differences in the refractive indices of running buffer and analyte solution, or unspecific binding of analyte to the chip. These effects are measured in a control experiment performed without receptor, and subtracted. Differences of a few millimolar in salt concentrations between buffers, or just 1% organic solvent from an analyte stock solution give rise to a strong background SPR signal.

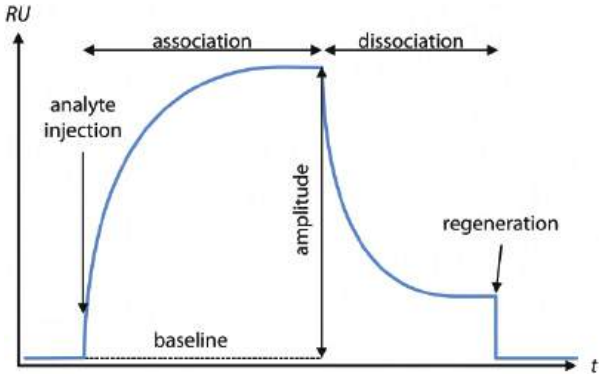


FIGURE 29.32: SPR sensorgram. The sensorgram is the change in SPR angle in resonance units (RU) as a function of time (blue curve). A buffer control is subtracted from the binding curve for analysis of association and dissociation phases. At the end of the experiment, unspecifically bound material is washed away (regeneration).

The association and dissociation phases of the sensorgram can be analyzed separately. From the association phase, two important values can be extracted: the amplitude, *i.e.* the total signal change relative to the baseline, and an observed association rate constant k_{obs} that is calculated by fitting a single exponential function to the data. Both amplitude and observed rate constant depend on the analyte concentration. For a reversible one-step binding reaction, the microscopic association rate constants k_{on} and k_{off} can be obtained from the linear dependence of k_{obs} on the analyte concentration (see Chapter 11). The dissociation rate constant k_{off} is directly obtained from the dissociation phase of the sensorgram. The dissociation of the complex is concentration-independent, and sensorgrams at different analyte concentrations should therefore yield the same k_{off} values. The ratio of k_{off} and k_{on} then yields a kinetically determined equilibrium dissociation constant K_d (eq. 29.48).

$$K_d = \frac{k_{off}}{k_{on}} \qquad \text{eq. 29.48}$$

(see eq. 8.21). In order for K_d to represent the binding equilibrium in solution, the surface-immobilized receptor molecules must behave the same way as in solution. Although K_d values determined by SPR are often comparable to the values determined in solution, fixing of the receptor to the sensor surface may change the binding entropy ΔS. In these cases, the equilibrium constant from SPR will be different from the K_d value measured in solution, and has to be considered as an apparent K_d value. These apparent K_d values still have qualitative value and can be compared within a series of experiments performed under identical conditions.

An alternative method estimates K_d from the amplitudes of sensorgrams measured at different analyte concentrations (Figure 29.33). Under non-saturating conditions, the sensor signal approaches a plateau termed the *equilibrium response*, where analyte binding and dissociation are in equilibrium.

The maximum or *saturation response* of the binding equilibrium is obtained under saturating conditions, *i.e.* with the analyte concentration much larger than the dissociation constant. Under these conditions, any receptor-analyte complex that dissociates will immediately be re-established by the sheer amount of analyte in the running buffer. The response units as a function of analyte concentration therefore provide a binding curve from which K_d can be determined. Global fitting procedures simultaneously evaluate a whole set of sensorgrams to derive k_{on}, k_{off}, and K_d from all curves.

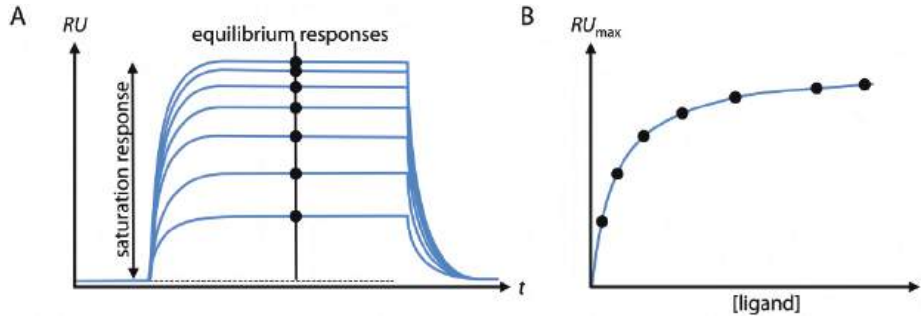


FIGURE 29.33: Binding constant from a set of sensorgrams. A: Several sensorgrams on top of each other with fast association and dissociation phases. B: Plot of amplitudes as a function of analyte concentration to yield K_d. This K_d value should be the same as the K_d derived from the kinetic constants.

SPR is suitable for the determination of association rate constants $< 10^5$ M^{-1} s^{-1}, and dissociation rate constants $> 10^{-5}$ s^{-1}. If k_{on} exceeds 10^5 M^{-1} s^{-1}, which is often the case for protein-protein interactions, there may be too few data points in the association phase for a reliable fit. Furthermore, when analyte binding is rapid, the analyte is depleted near the binding surface and its local concentration is lower than the concentration in the bulk solution, leading to errors in the analysis (Section 29.3.3). For slow dissociation reactions with k_{off} values below 10^{-5} s^{-1} (equivalent to a half-life of ca. 19 h), the dissociation phase cannot be recorded fully within a sensible time frame, and k_{off} has to be estimated from the initial slope of the dissociation curve and the expected amplitude. These limitations make SPR suitable for the kinetic determination of K_d values in the nanomolar range.

In principle, SPR experiments provide a set of thermodynamic values for binding. ΔG^0 can be calculated from the determined K_d. The enthalpy change ΔH can be obtained from the temperature dependence of K_d and a van't Hoff plot (Section 2.1.1), and the entropy change ΔS can be calculated from the Gibbs-Helmholtz equation (see Section 1.6; eq. 1.111). However, the effect of receptor immobilization on thermodynamic parameters should always be tested.

29.3.3 Mass transport limitation

When an analyte binds to its immobilized receptor during the association phase of the SPR experiment, the liquid layer close to the sensor chip will be depleted of analyte. Normally, convection and diffusion quickly replenish the depleted layer with analyte. If, however, the receptor density on the chip is very high and/or binding is very rapid, neither of these mechanisms can keep up with analyte depletion. The actual analyte concentration at the sensor chip will consistently be smaller than that in the flowing solution, a situation termed *mass transport limitation*. As a result, the observed association rate constants will also be smaller than their true values. A similar effect is observed during the dissociation phase. Any analyte dissociating from a complex can either be washed away or rebinds to the receptor. When receptor densities are high and/or the flow rate is low, re-binding is favored, and the observed dissociation rate constant is smaller than the true dissociation rate constant. One way to reduce the effects of mass transport is to increase the flow rate (from typically 10 µL min^{-1} to 100 µL min^{-1}), which will deliver analyte faster during binding or wash it away instantaneously during dissociation. A second way to reduce or sometimes entirely avoid mass transport limitation is to reduce the receptor density on the sensor chip by changing the immobilization conditions. This includes varying the receptor concentration, pH, ionic strength, and the activation and contact times.

29.3.4 Receptor immobilization on the sensor surface

In order to couple the receptor to the sensor surface, the gold layer first needs to be functionalized. This is achieved by coupling of a matrix that carries functional groups to the gold layer. Popular matrices include cellulose, agarose, alginate, polycarboxylate hydrogel, and, in particular, carboxymethyl-functionalized (CM-)dextran. By using dextran with carboxymethyl groups in various densities, different receptor densities can be achieved. The carboxymethyl groups are then activated and used for introduction of other functionalities to which the receptor is bound covalently or non-covalently.

29.3.4.1 Covalent receptor immobilization

Amine coupling is the most frequently used technique for direct immobilization of receptors. The carboxyl groups of the CM-dextran matrix are activated by a carbodiimide (Figure 29.34) and converted into N-hydroxysuccinimide (NHS) esters. These esters then react with amine nucleophiles of the receptor such as the abundant Lys side chains or the N-terminus. The result is a stable amide bond of the receptor with the sensor surface. Unreacted functional groups are blocked with an excess of ethanolamine.

Due to the high abundance of lysine in proteins and the large number of amino groups, amine coupling may lead to a wide spectrum of different receptor orientations relative to the surface, and not all of the receptors may be capable of binding to their binding partner.

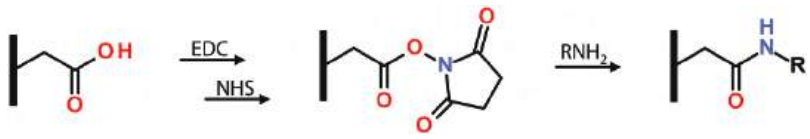


FIGURE 29.34: Amine coupling. Carboxylate groups are activated, for instance, by the soluble 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC). Substitution with N-hydroxysuccinimide (NHS) gives a reactive ester and the urea derivative of EDC. Substitution of the NHS-ester by an amine-containing receptor couples the receptor to the surface *via* an amide bond. Ethanolamine is then used in excess to block unreacted esters.

Thiol coupling uses disulfide exchange reactions to attach a receptor containing free thiol groups to the sensor chip *via* a disulfide bond (Figure 29.35). A disulfide group in the matrix is generated by activating the carboxylate groups of the dextran with EDC and NHS just as in the amine coupling procedure, and the resulting NHS ester is reacted with a disulfide-containing amine that contains a good leaving group. The receptor is then coupled to the surface by disulfide exchange.

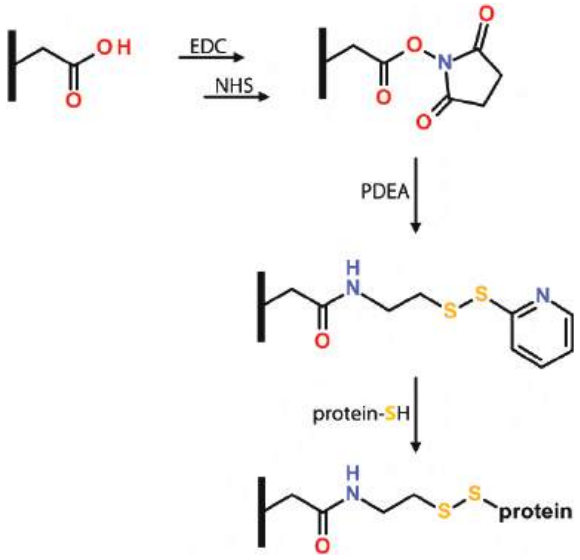


FIGURE 29.35: Thiol coupling. 2-(2-pyridinyldithio) ethaneamine (PDEA) is amine-coupled to the dextran matrix, followed by disulfide exchange with the receptor. 2-thiopyridine is a good leaving group. Excess free Cys is used to block unreacted PDEA groups.

Cysteines in the receptor can be used for thiol-coupling. Because of the lower abundance of cysteines compared to lysines, thiol coupling leads to more uniform orientations of the receptor on the surface than amine coupling. If the receptor lacks cysteines, they can be introduced by site-directed mutagenesis.

Maleimide coupling is also based on thiol chemistry. Here, Cys residues in proteins react with a matrix derivatized with maleimide groups (Figure 29.36). The resulting esters are stable in basic conditions and insensitive to reducing agents that may be required in the SPR buffer.

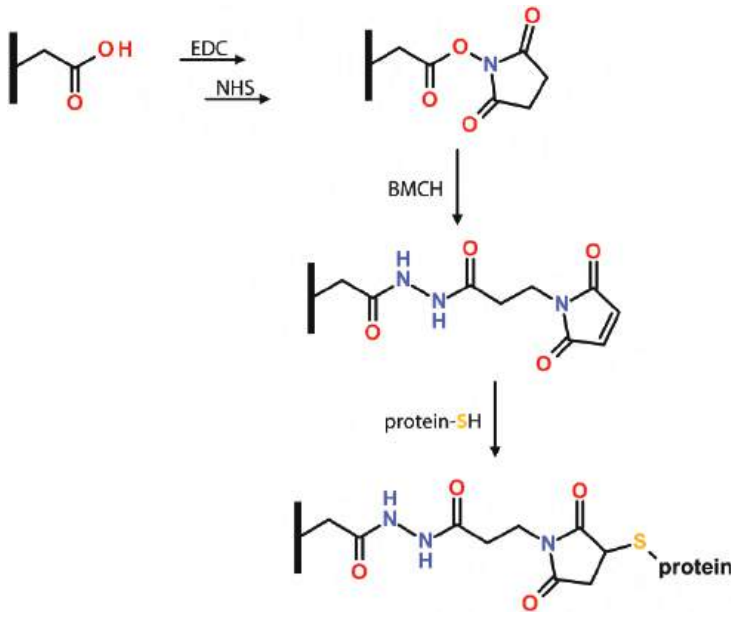


FIGURE 29.36: Maleimide coupling. Any reactive maleimide can be used to substitute the NHS-ester, for instance BMCH (N-[β-maleimidopropionic acid]-hydrazide), or EMCH (N-[ε-maleimidocaproic acid]-hydrazide), which has a longer alkyl linker.

Aldehyde coupling (Figure 29.37) is useful for immobilizing glycoproteins *via* their carbohydrate moieties on hydrazide-functionalized surfaces. If the carbohydrate possesses a reducing end, this aldehyde group can be directly used for coupling. If not, a *cis*-diol in the carbohydrate is oxidized to form aldehyde groups. The aldehyde is then condensed with the hydrazide on the surface to form a hydrazone, which is further reduced to a stable amine. As carbohydrates occur at specific sites on the protein surface (Section 18.1.4.1) carbohydrate coupling leads to homogenously immobilized receptors on the sensor surface.

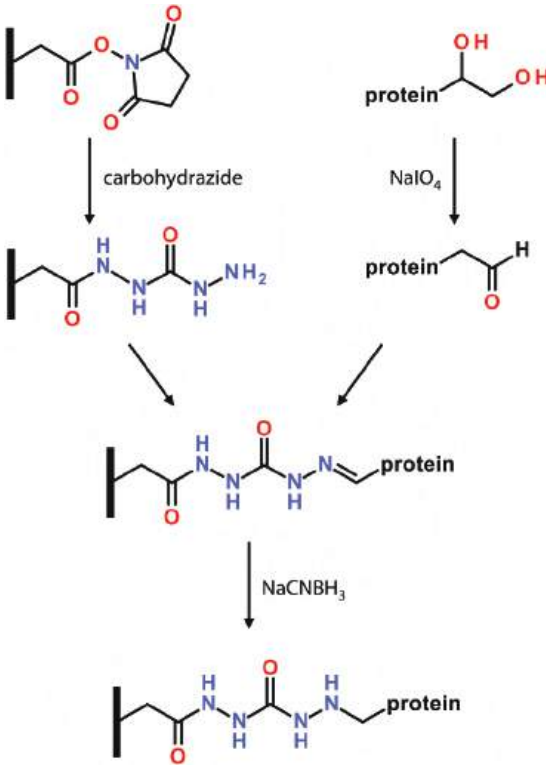


FIGURE 29.37: Aldehyde coupling. The matrix is functionalized by reaction of the NHS ester with carbohydrazide. The receptor aldehyde groups, generated by oxidation of *cis*-diols if necessary, are condensed with the hydrazide to a hydrazone that is reduced to a stable amine by $NaCNBH_3$.

Apart from these organic chemical coupling reactions, biochemical coupling of modified peptides to the receptor using intein-mediated protein ligation or the transpeptidase *sortase* is also possible (see Section 21.5.4).

29.3.4.2 Non-covalent receptor immobilization

Non-covalent immobilization uses the interaction of surface-coupled binding modules such as monoclonal antibodies, metal chelators, or streptavidin with affinity tags on the receptor. The binding module is covalently coupled to the surface by the same techniques used for covalent receptor coupling (Section 29.3.4.1). The receptor is produced with an affinity tag that is recognized by the binding module, such as the epitope for an antibody, a His-tag, or a biotin group, and coupled to the surface. Non-covalent coupling allows complete regeneration of the sensor chip and reuse with the same or a different receptor, which is impossible with covalently attached receptors.

Non-covalent immobilization uses the interaction with a unique moiety on the receptor, and usually leads to a rather homogeneous orientation of the receptor on the sensor. The affinity of the interaction between the affinity tag on the receptor and the binding module on the surface should be high enough to prevent receptor dissociation during the experiment, which leads to a baseline drift.

Proteins can be chemically biotinylated at Lys residues using biotinyl-N-hydroxysuccinimide esters. Site-specific biotinylation can be achieved by expressed protein ligation of the receptor with a biotinylated peptide, by introduction of a biotinylation sequence and *in vivo* or *in vitro* biotinylation at a specific lysine in this sequence with biotin ligase, or by incorporation of a biotinylated amino acid through *in vitro* translation of mRNA and stop-codon suppression (Section 18.1.5; see also Section 21.5.4.2). Biotinylated proteins are then immobilized on streptavidin-functionalized surfaces. The dissociation rate of the streptavidin-biotin complex is low (k_{off} ca. 10^{-6} s^{-1}), leading to negligible baseline drifts during experiments that may take several hours to complete.

Monoclonal antibodies can be covalently coupled to the sensor chip. The epitope recognized by the antibody, such as glutathione S transferase (GST), the His-tag, or the FLAG tag (see also Section 21.5.4.2), is fused to the receptor, which is then directly coupled to the surface by antibody-antigen interactions.

Metal chelation is a standard affinity chromatographic step during protein purification. The same principle can be applied for coupling of receptors to SPR sensor chips. The dextran matrix of the sensor chip is coupled to nitrilotriacetic acid (NTA) that chelates divalent cations (mostly Ni^{2+}, but Zn^{2+} and Co^{2+} are also used). The protein of interest carries an N- or C-terminal His-tag, which binds to the NTA-metal complex.

29.3.5 Stoichiometry of binding in an SPR experiment

SPR experiments can be used to determine binding stoichiometries. The maximum amount of analyte that can be bound depends on the receptor density RU_{receptor} (Section 29.3.3) and on the fraction P of the receptor that is active, *i.e.* which is immobilized in an orientation that does not occlude the binding site. With random chemical coupling, the fraction P of active receptor can be as low as 0.2. P can be determined by measuring the maximal SPR response $RU_{\text{known analyte,max}}$ by an analyte that binds to the receptor with known stoichiometry N_{known}:

$$P = \frac{1}{N_{\text{known}}} \cdot \frac{RU_{\text{known analyte, max}}}{RU_{\text{receptor}}} \cdot \frac{M_{\text{M,receptor}}}{M_{\text{M,known analyte}}} \qquad \text{eq. 29.49}$$

Eq. 29.49 takes into account that the measured response only depends on the molecular masses M_M of analyte and receptor. Once P is known for the reference system, the stoichiometry for an unknown analyte binding to the same sensor chip coupled with this receptor can be calculated from the maximum signal $RU_{\text{unknown analyte,max}}$ obtained by the unknown analyte.

$$N = \frac{RU_{\text{unknown analyte, max}}}{P \cdot RU_{\text{receptor}}} \cdot \frac{M_{\text{M,receptor}}}{M_{\text{M,unknown analyte}}} \qquad \text{eq. 29.50}$$

The stoichiometry of binding for an unknown analyte can thus be obtained in comparison to a reference measurement. Stoichiometries deviating from the expected number of binding sites may point to unspecific binding of the analyte to the sensor surface. Stoichiometries larger than the number of binding sites (super-stoichiometric binding) are often observed in the analysis of weak interactions, where the analyte concentrations have to be very high (Box 29.2).

29.3.6 Specificity of binding in an SPR experiment

A response signal in SPR experiment reports on any mass deposition on the sensor surface, and may originate either from specific binding of the analyte to the receptor binding site or from unspecific binding to the surface of the receptor or directly to the sensor. Specific and unspecific binding can be distinguished in control experiments. To monitor unspecific binding to the sensor surface, a reference channel is used that has no receptor immobilized. This type of control experiment can be used to minimize unspecific binding by optimization of buffer composition, temperature, and flow rate. Unspecific binding to the immobilized receptor outside the binding site can be estimated in a control experiment with an inactive receptor. This inactive receptor may have an amino acid substitution in the binding site that interferes with binding of the analyte, or the binding site may be blocked by modification by a covalent inhibitor. If an SPR response is obtained with such a blocked receptor, the signal is likely due to unspecific binding. Careful controls for unspecific binding are particularly important in the validation of binding events during small molecule screening with SPR (Box 29.2).

BOX 29.2: USING SPR IN SMALL MOLECULE SCREENING.

SPR can be automated to test binding of several thousand compounds to a receptor within a matter of days or a few weeks. In SPR-based screening of small molecule binding to macromolecular drug targets, the drug target is coupled to the surface as the receptor, and a small molecule library is tested for binding at a single analyte concentration. The SPR response scales with the molecular mass of the compound, and testing the binding of small molecules (M_M 100–250 Da) to a macromolecule brings SPR near its current detection limit (ca. 50 Da or a few *RU*). To maximize the measured signal, high densities of receptor and high fractions of active receptors are needed. The resulting mass transport limitation is acceptable as long as no kinetic information is desired. High receptor densities are affordable because the overall receptor consumption in SPR is small, typically less than 100 µg per screen. Small molecules usually bind with low affinity to the receptor, with K_d values in the micromolar to millimolar range, and SPR experiments therefore require high concentrations of analyte (10–200 µM). Unspecific binding that leads to false positive hits therefore needs to be accounted for by suitable controls (Section 29.3.6). Molecules that elicit an SPR response (*positive hits*) are validated in a *dose-response* or titration experiment to obtain binding curves and to possibly gain insight into the binding kinetics. Further validation of binding uses an orthogonal technique such as NMR or structure determination by X-ray crystallography. Using SPR-based small molecule screening, starting molecules for drug design have been identified for a number of pharmaceutical targets, including HIV-1 protease and HIV-1 reverse transcriptase, but also for the G-protein coupled receptors A_{2A} and CCR5.

SPR is now a standard technique to characterize molecular interactions. SPR is particularly well suited to study ligand binding to membrane receptors, because the immobilization of membrane proteins on the surface emulates their two-dimensional arrangement in a lipid bilayer. A recent development is to solubilize membrane proteins in *nanodiscs* (Section 18.4.5), which are small patches of lipid bilayers where the hydrophobic rim is shielded from solvent by the *Membrane Scaffold Protein* (MSP). MSP is a derivative of the human apolipoprotein A-I, which forms ring-shaped dimers that encircle the nanodisc (Figure 18.50). The whole nanodisc, including the dissolved membrane protein and MSP, is then attached to a sensor chip. SPR has also been used to detect conformational changes in proteins by way of changes in mass distribution of surface-immobilized proteins.

QUESTIONS

29.1 A rotor spins at 60000 rpm. What is the radial position at which a centrifugal field of 250000·*g* is generated? What would be the apparent mass of an imbalance of 100 mg at this position?

29.2 You want to study the assembly of a nucleotide-binding protein by AUC. In order for the protein to form a complex, high concentrations of nucleotide are required. What detection methods are best suited? Another protein binds very tightly to its nucleotide, and upon binding of the nucleotide it forms oligomers at very low (nanomolar) concentrations. What probe located at which binding partner would you choose to detect such an assembly?

29.3 Using l'Hôpital's rule, show that for an oblate spheroid:

$$\lim_{a\to b}\frac{f}{f_0}=\lim_{a\to b}\frac{\sqrt{\frac{b^2}{a^2}-1}}{\sqrt[3]{\frac{b^2}{a^2}}\cdot\arctan\left(\sqrt{\frac{b^2}{a^2}-1}\right)}=1$$

The derivative of arc tan (x) is:

$$\frac{d}{dx}\arctan(x)=\frac{1}{x^2+1}$$

29.4 Using l'Hôpital's rule, show that for a prolate spheroid:

$$\lim_{a\to b}\frac{f}{f_0}=\lim_{a\to b}\frac{\sqrt{\frac{a^2}{b^2}-1}}{\sqrt[3]{\frac{a}{b}}\cdot\ln\left(\frac{a}{b}+\sqrt{\frac{a^2}{b^2}-1}\right)}=1$$

29.5 Given is the Kirkwood approximation,

$$f_{\mathrm{N}}=Nf_1\left(1+\frac{f_1}{6\pi\eta N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{1}{R_{\mathrm{ij}}}\right)^{-1} \tag{1}$$

and the relation of molar mass and friction coefficient to the sedimentation coefficient,

$$s=\frac{M\left(1-\bar{\nu}\rho\right)}{N_{\mathrm{A}}\cdot f} \tag{2}$$

and Stokes' law:

$$f_1=6\pi\eta R_{\mathrm{e}} \tag{3}$$

show that for a homo-oligomer the following relation holds:

$$\frac{s_{\mathrm{N}}}{s_1}=1+\frac{R_{\mathrm{e}}}{N}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{1}{R_{\mathrm{ij}}} \tag{4}$$

where R_{e} is the Stokes radius of the monomer with friction coefficient f_1, and s_{N} and s_1 are the sedimentation coefficients of the oligomer and the monomer, respectively.

29.6 Using the result from the previous question, show that for a dimer (N = 2) and trimer (N = 3) of spheres, the expected ratios s_{N}/s_1 are 1.25 and 1.5, respectively.

29.7 A tetramer sediments with a coefficient of s_4 = 33.5 S. The monomer sediments with s_1 = 20 S. Is the tetramer square planar or tetrahedral?

29.8 Starting from this form of the Lamm equation:

$$\frac{\partial c}{\partial t}=\frac{1}{r}\cdot\frac{\partial}{\partial r}\left(r\cdot D\cdot\frac{\partial c}{\partial r}-s\omega^2r^2c\right)$$

derive the Lamm equation with the separate contributions of diffusion and sedimentation:

$$\frac{\partial c}{\partial t} = D \cdot \left(\frac{\partial c}{r \cdot \partial r} + \frac{\partial^2 c}{\partial r^2} \right) - s\omega^2 \cdot \left(2c + r \cdot \frac{\partial c}{\partial r} \right)$$

29.9 How much (in %) does the observed s value for a globular protein at 10 mg mL^{-1} differ from that extrapolated to zero concentration (k_s = 0.005 mL mg^{-1})? What concentration would be needed to have only a 0.1% change of s?

29.10 A 50 kDa protein is sedimented under equilibrium conditions at 300 K in a buffer of density 1 kg·L^{-1}. Calculate the rotation speed required to produce a concentration ratio of 100 between the meniscus (r_0 = 6.2 cm) and a radial position 5 mm downwards of the meniscus. The partial specific volume can be assumed as $\bar{\nu}$ = 0.73 L·kg^{-1}. The concentration ratio is given by

$$\ln \frac{c(r)}{c_0} = \frac{M(1-\bar{\nu}\rho)\omega^2\left(r^2 - r_0^2\right)}{2RT}$$

29.11 Using the formula for the radius of gyration

$$R_g = \sqrt{\frac{\sum m_i r_i^2}{\sum m_i}}$$

show that R_g for a sphere of radius R equals $R_g = \sqrt{0.4} \cdot R$.

29.12 The radius of gyration of DNA can be approximated by that of a solid cylinder, rotated about its centroid axis:

$$R_g^2 = \frac{R^2}{4} + \frac{L^2}{12}$$

What would R_g be for a double-stranded B-DNA of 100 base pairs? Is it very different from a sphere where the dsDNA would fit into?

29.13 For r_0 = 1 cm, z_0 = 0.783 cm, U_1 = 1000 V, and ω = 1.1 MHz, what is the minimum m/z that can be stored in the ion trap of a mass spectrometer? The Avogadro constant is 6.022·10^{23} mol^{-1} and 1 V = 1 kg m^2 s^{-2} C^{-1}.

29.14 In a mass spectrometer we would like to measure the current of an ion with a single positive charge for as long as it takes to reach a signal-to-noise ratio of ten. The noise level of the detector is 1 fA and we assume a counting efficiency of unity. The ion current is 300 Hz. How long do we have to measure? Hint: For counting processes (Poisson statistics), the signal-to-noise ratio increases with the square root of the integration time:

$$\frac{S}{N} \sim \sqrt{t}$$

29.15 To circumvent the long integration time from the previous question, we use a photomultiplier with a gain of 10^5 to boost the signal. How long would we need to measure for S/N = 10, assuming a counting efficiency of unity?

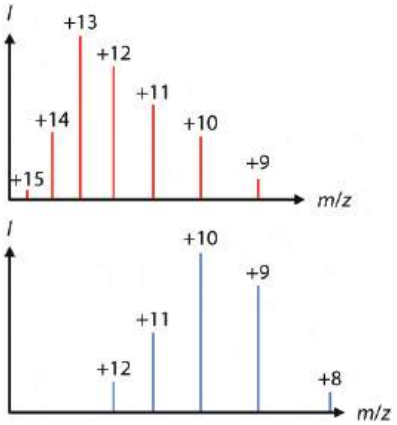
29.16 The balance group in the iTRAQ reagent has only two atoms, carbon and oxygen. Which isotope combinations are used to achieve the four different masses of 28–31 Da?

29.17 The time of flight of an ion after MALDI is 0.2 ms. The acceleration voltage is 15 kV, and the length of the drift tube is 2 m. Only a single ion type is observed. What is the molecular mass of the ion in Da? Is it a peptide? The elemental charge is $1.602 \cdot 18 \cdot 10^{-19}$ C and 1 Da equates to $1.6605 \cdot 10^{-27}$ kg. What would be the time of flight for a peptide of molecular mass 1000 Da under these conditions? Would the same peptide with two charges take half the time?

29.18 Se-Met modified proteins are used for phasing in X-ray crystallography, and ^{15}N-labeled proteins are used for 2D NMR spectroscopy. How can the percentage of incorporation of these modifications be measured by mass spectrometry?

29.19 What is the expected relative mass peak distribution for ethane? What are the first three relative peak intensities for fullerene C_{60}? The natural abundances of ^{12}C and ^{13}C are 98.9% and 1.1%, respectively.

29.20 A protein is analyzed by native and denaturing ESI. The following spectra are obtained:



Which of the spectra was recorded for the native protein and why is the ionization pattern different?

29.21 Lysozyme from chicken has a calculated molecular mass of 16238.6 Da. The sequence contains eight cysteine residues. ESI mass analysis gives the following *m/z* data:

m/z
2706.08
2319.68
2029.81
1804.40
1624.04
1476.53
1353.57

Assign the correct charges to the *m/z* data and calculate the mean of the mass. What could explain the mass difference to the calculated value?

29.22 Why can samples be colored, turbid, or even opaque, or why does absorption or scattering not pose a problem for SPR?

29.23 To reduce mass transport during the dissociation phase in an SPR experiment, someone suggests injecting an inhibitor for either the receptor or the analyte in order to avoid rebinding. What conditions must such an inhibitor meet?

29.24 Why is Tris buffer not suitable for amine coupling?

29.25 Thiol coupling is often used for acidic proteins with a pI of < 3.5. Why could that be?

29.26 In some SPR instruments, one *response unit* or *resonance unit* (*RU*) corresponds to 0.0001° shift in SPR angle, which is equivalent to a refractive index change of 10^{-6} and a mass change on the sensor surface of about 1 pg mm^{-2}. A receptor of molecular mass 50 kDa (for example, an antibody Fab) is immobilized on an SPR chip, which leads to an increase in the response by 1000 *RU*. What are the changes in resonance angle and refractive index? What are the weight and molar receptor densities? A ligand of molecular mass 500 Da that is known to bind in a 1:1 stoichiometry to the Fab is passed over the chip and elicits an additional response signal of 5 *RU*. What are the expected signal and, consequently, the fraction of active receptor?

29.27 The equilibrium responses in an SPR experiment can be used to derive a K_d value for a binding equilibrium A + B ⇄ AB. Since the total concentration of the partner bound to the sensor chip, $[A]_0$, is not known exactly, the degree of binding $\nu = [AB]/[A]_0$ is used instead. Derive the fitting equation for this experiment, a variant of the so-called Langmuir equation.

29.28 What is the error propagation for ΔG when calculated from an equilibrium dissociation constant K_d that was itself calculated from kinetic constants k_{on} and k_{off} (and their errors)? Is the unit of the error on ΔG as expected? What would the error for ΔG calculate to if the errors on k_{on} and k_{off} were 10% of the measured values?

29.29 Show that the error propagation for an equilibrium dissociation constant K_d when obtained from kinetic constants k_{on} and k_{off} is (should be) identical compared to the error derived from a measured equilibrium association constant K_a (*e.g.* from an ITC experiment).

REFERENCES

Bocquet, N., Kohler, J., Hug, M. N., Kusznir, E. A., Rufer, A. C., … Huber, S. (2015) Real-time monitoring of binding events on a thermostabilized human A2A receptor embedded in a lipid bilayer by surface plasmon resonance. *Biochim Biophys Acta* **1848**(5): 1224–1233.
· ligand binding to a GPCR embedded in nanodiscs is studied by SPR

Boussaad, S., Pean, J. and Tao, N. J. (2000) High-resolution multiwavelength surface plasmon resonance spectroscopy for probing conformational and electronic changes in redox proteins. *Anal Chem* **72**(1): 222–226.
· detection of conformational changes in proteins by SPR

Clow, F., Fraser, J. D. and Proft, T. (2008) Immobilization of proteins to biacore sensor chips using *Staphylococcus aureus* sortase A. *Biotechnol Lett* **30**(9): 1603–1607.
· protein coupling to surfaces using sortase

Fletcher, J. S., Lockyer, N. P., Vaidyanathan, S. and Vickerman, J. C. (2007) TOF-SIMS 3D biomolecular imaging of *Xenopus laevis* oocytes using buckminsterfullerene (C60) primary ions. *Anal Chem* **79**(6): 2199–2206.
· C_{60^+} ions were used to gently etch the surface of *Xenopus* oocytes at ambient temperature and reconstruct a 3D image of triacyl glyceride distribution

Fletcher, J. S., Vickerman, J. C. and Winograd, N. (2011) Label free biochemical 2D and 3D imaging using secondary ion mass spectrometry. *Curr Opin Chem Biol* **15**(5): 733–740.
· description of the SIMS method

Glish, G. L. and Vachet, R. W. (2003) The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* **2**(2): 140–150.
· review of mass spectrometry

Hall, D. R., Cann, J. R. and Winzor, D. J. (1996) Demonstration of an upper limit to the range of association rate constants amenable to study by biosensor technology based on surface plasmon resonance. *Anal Biochem* **235**(2): 175–184.
· discusses how mass transport limitation puts an upper limit to the association rate constants measurable by SPR

Huber, W. and Mueller, F. (2006) Biomolecular interaction analysis in drug discovery using surface plasmon resonance technology. *Curr Pharm Des* **12**(31): 3999–4021.

Navratilova, I. and Hopkins, A. L. (2011) Emerging role of surface plasmon resonance in fragment-based drug discovery. *Future Med Chem* **3**(14): 1809–1820.
· reviews of the physics and applications of SPR with an emphasis on drug design

Klatt, S., Hartl, D., Fauler, B., Gagoski, D., Castro-Obregon, S. and Konthur, Z. (2013) Generation and characterization of a Leishmania tarentolae strain for site-directed *in vivo* biotinylation of recombinant proteins. *J Proteome Res* **12**(12): 5512–5519.
· use of biotin ligase for biotinylation

Kramer, K., Sachsenberg, T., Beckmann, B. M., Qamar, S., Boon, K. L., Hentze, M. W., Kohlbacher, O. and Urlaub, H. (2014) Photo-crosslinking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat Methods* **11**(10): 1064–1070.
· identifying protein-RNA interaction sites by photo-crosslinking and mass-spectrometric analysis

Levary, D. A., Parthasarathy, R., Boder, E. T. and Ackerman, M. E. (2011) Protein-protein fusion catalyzed by sortase A. *PLoS One* **6**(4): e18342.
· protein-protein coupling using sortase

Lue, R. Y., Chen, G. Y., Hu, Y., Zhu, Q. and Yao, S. Q. (2004) Versatile protein biotinylation strategies for potential high-throughput proteomics. *J Am Chem Soc* **126**(4): 1055–1062.
· intein-mediated biotinylation of proteins

Meselson, M. and Stahl, F. W. (1958) The replication of DNA in *Escherichia coli*. *Proc Natl Acad Sci U S A* **44**(7): 671–682.
· original description of the Meselson-Stahl experiment

Navratilova, I., Papalia, G. A., Rich, R. L., Bedinger, D., Brophy, S., … Myszka, D. G. (2007) Thermodynamic benchmark study using Biacore technology. *Anal Biochem* **364**(1): 67–77.
· the thermodynamic parameters K_d, ΔH, and ΔS of carbonic anhydrase II measured by SPR and ITC were in agreement, indicating that fixing of the enzyme on the sensor surface does not strongly alter the entropy

Ostrowski, S. G., Van Bell, C. T., Winograd, N. and Ewing, A. G. (2004) Mass spectrometric imaging of highly curved membranes during *Tetrahymena* mating. *Science* **305**(5680): 71–73.
· dramatic change in lipid composition during mating of *Tetrahymena* detected by MS

Perspicace, S., Banner, D., Benz, J., Muller, F., Schlatter, D. and Huber, W. (2009) Fragment-based screening using surface plasmon resonance technology. *J Biomol Screen* **14**(4): 337–349.
· small molecule screening of a library identifies inhibitors for the serine protease chymase

Sota, H., Hasegawa, Y. and Iwakura, M. (1998) Detection of conformational changes in an immobilized protein using surface plasmon resonance. *Anal Chem* **70**(10): 2019–2024.
· detection of conformational changes in proteins by SPR

Tsao, K. L., DeBarbieri, B., Michel, H. and Waugh, D. S. (1996) A versatile plasmid expression vector for the production of biotinylated proteins by site-specific, enzymatic modification in *Escherichia coli*. *Gene* **169**(1): 59–64.
· use of biotin ligase for biotinylation

ONLINE RESOURCES

www.analyticalultracentrifugation.com
· description of the Sedfit program

leonardo.inf.um.es/macromol/programs/programs.htm
· description of the Hydropro program

masspec.scripps.edu/book_toc.php

www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm
· MS tutorials

www.sprpages.nl
· background information on SPR

30 Calorimetry

Calorimetric methods measure the heat released (exothermic process) or taken up (endothermic process) during a reaction and are among the few methods that directly yield the associated changes in thermodynamic functions including ΔH, ΔG, ΔS, and ΔC_p. There are two implementations of calorimetry, isothermal titration calorimetry (ITC) and differential scanning calorimetry (DSC). ITC (Section 30.1) provides thermodynamic information on binding events and chemical reactions, whereas DSC (Section 30.2) finds applications in the thermodynamic characterization of phase transitions.

30.1 ISOTHERMAL TITRATION CALORIMETRY

30.1.1 General principle

In isothermal titration calorimetry (ITC), heat is used as a direct probe to follow binding reactions and to obtain binding curves. An isothermal titration calorimeter (Figure 30.1) consists of two cells, one for the sample and one for the reference buffer. The cells are made of a non-corroding metal alloy and are thermally insulated from the surroundings by an *adiabatic jacket*. The sample cell contains one of the binding partners in a suitable buffer. A syringe that holds the second binding partner in the same buffer is inserted into the sample cell. During an isothermal titration, this binding partner is injected in small steps into the sample cell. The syringe rotates throughout the experiment, and a paddle at its end acts as a stirrer to ensure rapid mixing. The reference cell contains the buffer in which the binding partners are dissolved.

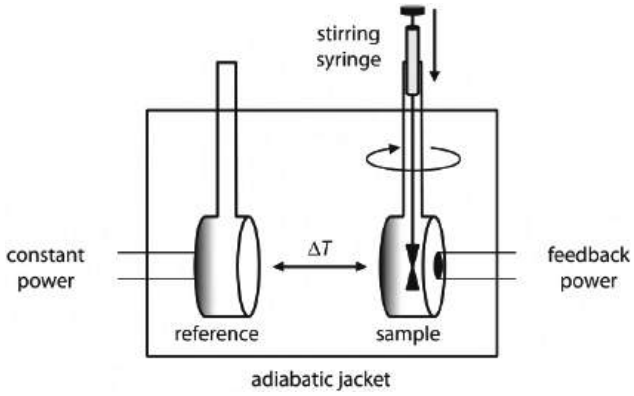


FIGURE 30.1: Principle of an ITC. In ITC, the temperature of the reference cell is held constant (at the chosen temperature of the experiment) and the feedback power is used to compensate for any change in heat released or taken up by the reaction in the measurement cell. The adiabatic jacket keeps the whole setup at the temperature of the experiment. The paddle at the end of the syringe holding the second binding partner serves as a stirrer. Injection of a certain volume from the syringe (usually 3–15 μL at a rate of 0.5 μL s^{-1}) displaces the same volume from the measurement cell. This loss of material has to be taken into account during data analysis.

The primary readout for heat in modern ITC (and DSC) instruments is electric power. Electrical energy (in the form of a heating current) is used to keep the reference cell and the sample cell at a constant temperature difference ΔT. The sample cell is kept at a slightly higher temperature than the reference cell by a positive *feedback power*. Maintaining a constant temperature difference between the cells, instead of keeping both cells at the same temperature, offers the advantage that the feedback system can immediately respond to a temperature increase and to a temperature decrease in the reaction cell, simply by adjusting the feedback power. In exothermic reactions, heat is released in the sample cell in each titration step, leading to an increase in temperature. To maintain the constant temperature difference to the reference cell, the feedback power to the sample cell is reduced. For endothermic reactions, on the other hand, the feedback power needs to be increased to compensate for heat taken up. In either case the feedback power remains positive. It is important to equilibrate the entire system at the beginning, and to not start the first injection before a stable baseline of feedback power is recorded. Similarly, after each titration step the baseline has to be reached again before the next injection can follow. In principle, the sample and reference cells could also be maintained at the same temperature throughout the experiment. In this case, a temperature decrease in the sample cell during endothermic reactions would have to be compensated by cooling of the reference cell instead of heating of the sample cell. However, cooling is much slower than the instantaneous reduction in feedback power. Accordingly, the response time of a calorimeter operated in this *T-difference mode* is smaller (a few seconds) than the response time of a calorimeter operating at constant temperature. ITC experiments in the *T*-difference mode even allow measurement of slow enzymatic reactions (Section 30.1.6).

The raw data from an isothermal titration is a time trace of the feedback power, with a negative peak for each titration step for exothermic reactions (less feedback power required; Figure 30.2), and a positive peak for each step for endothermic reactions (more feedback power required; Figure 30.2 would appear flipped about the horizontal axis). Power equals energy (or work) per time, and has the unit Watt, with 1 W = 1 J s^{-1}. Integration of the electric power over time therefore directly provides the heat exchanged. Before the peaks can be integrated, the raw titration data need to be corrected for a general drift of the instrument baseline. Since introduction of the first titration calorimeter in 1989, ITC instruments have become more and more sensitive, and modern instruments measure heat changes in the μJ range.

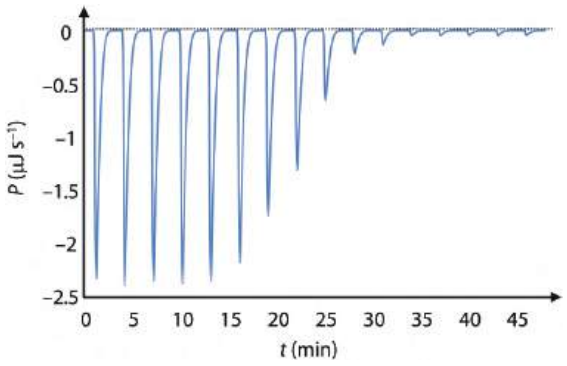


FIGURE 30.2: Schematic raw ITC data for an exothermic reaction. After correction for the instrument baseline drift, the baseline for the titration is horizontal. Heat released in each titration step is measured as a reduced feedback power. At the end of the experiment, only heat of dilution is measured. An endothermic reaction would give a signal flipped along the horizontal axis. The number of injections (10–50) and the equilibration time per injection (1–5 min) determine the total experiment time. For data analysis, each peak is integrated.

During data analysis, the heat exchange in each titration step is calculated by integration of the power curve over time for each injection. From the known concentrations of the binding partners in the sample cell and the syringe, the cell volume, and the volume of the injection, the concentration of both binding partners at each titration step is calculated. Plotting the heat per injection as a function of the molar ratio of the binding partners yields a sigmoidal binding curve also known as an *enthalpogram* (Figure 30.3). The enthalpogram can then be analyzed to obtain the binding enthalpy, ΔH, the equilibrium association constant K_a (typically, ITC data are evaluated as $K_a = 1/K_d$), and the binding stoichiometry n.

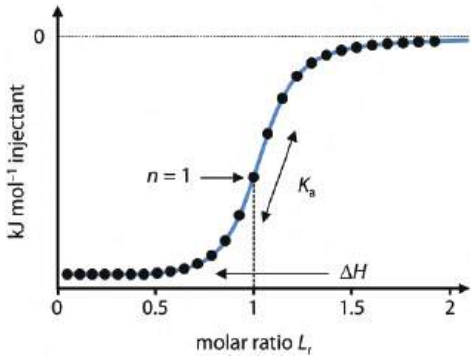


FIGURE 30.3: A binding curve resulting from integration of the raw ITC data. The shape of the curve allows conclusions on the thermodynamic parameters. The inflection point corresponds to the stoichiometry of binding n. The constant heat at the beginning of the titration corresponds to ΔH of the reaction. The steepness of the curve is related to the association constant K_a, the affinity.

30.1.2 ITC DATA ANALYSIS

For a simple 1:1 binding event of ligand L and macromolecule M, the association constant K_a is

$$K_a = \frac{[ML]}{[M]\cdot[L]} \qquad \text{eq. 30.1}$$

$[L]$, $[M]$, and $[ML]$ are the equilibrium concentrations of free ligand, free macromolecule, and complex, respectively. The starting concentration of ligand is L_0, and the starting concentration of the macromolecule is M_0. Substituting the mass conservation $[M] = M_0 - [ML]$ and $[L] = L_0 - [ML]$ into eq. 30.1 followed by rearrangement yields

$$[ML]^2 - \left(L_0 + M_0 + \frac{1}{K_a}\right) \cdot [ML] + 4M_0L_0 = 0 \qquad \text{eq. 30.2}$$

This quadratic equation has only one physically sensible solution (the other would yield ML concentrations exceeding M_0, which is impossible, see Section 2.2):

$$[ML] = \frac{1}{2} \cdot \left(L_0 + M_0 + \frac{1}{K_a} - \sqrt{\left(L_0 + M_0 + \frac{1}{K_a}\right)^2 - 4M_0L_0} \right) \qquad \text{eq. 30.3}$$

The heat we determine for each titration step of an ITC experiment is related to the additional formation of ML at the ligand concentration L in this step. We therefore differentiate eq. 30.3 with respect to the total ligand concentration L_0 and simplify to

$$\frac{d[ML]}{dL_0} = \frac{1}{2} \cdot \left(1 - \frac{L_0 - M_0 + \frac{1}{K_a}}{\sqrt{\left(L_0 + M_0 + \frac{1}{K_a}\right)^2 - 4M_0L_0}} \right) \qquad \text{eq. 30.4}$$

This equation can be further simplified by defining the molar ratio L_r of the binding partners

$$L_r = \frac{L_0}{M_0} \qquad \text{eq. 30.5}$$

and

$$r = \frac{1}{C} = \frac{1}{K_a M_0} \qquad \text{eq. 30.6}$$

C is the *Wiseman constant*. We will see later in this Section that the Wiseman constant helps us in designing ITC experiments. In order to substitute these helper variables into eq. 30.4 we multiply both numerator and denominator with $1/M_0$ and rearrange to

$$\frac{d[ML]}{dL_0} = \frac{1}{2} \cdot \left(1 - \frac{L_r - 1 + r}{\sqrt{L_r^2 - 2L_r(1-r) + (1+r)^2}} \right) \qquad \text{eq. 30.7}$$

We now have to relate the change in concentration of the ML complex in each titration step, d[ML], to the measured heat dq. The heat dq required for compensating the release or take-up of heat during the reaction depends on the amount of complex d(ML) that is formed, the reaction enthalpy ΔH, and the volume of the sample cell V_0:

$$dq = d[ML] \cdot \Delta H \cdot V_0 \qquad \text{eq. 30.8}$$

Upon combining eq. 30.7 and eq. 30.8 we arrive at the expression that describes our ITC binding curve (Figure 30.3):

$$\frac{dq}{dL_0} = \frac{\Delta H \cdot V_0}{2} \cdot \left(1 - \frac{L_r - 1 + r}{\sqrt{L_r^2 - 2L_r(1-r) + (1+r)^2}} \right) \qquad \text{eq. 30.9}$$

Non-linear least square fitting of eq. 30.9 to the ITC data then yields ΔH, the molar ratio of the binding partners L_r, and the association constant K_a (through the parameter r, see eq. 30.6). In practice, a

correction is needed to account for the volume that is displaced from the sample cell by injection of ligand from the syringe (Figure 30.1). To keep this correction small, it is necessary to limit injections to small volumes relative to the sample cell volume V_0. For injections of 3–15 µL into a 1.4 mL cell, the effect of displacement of solution from the sample cell can be neglected. Injection volumes can be kept small by placing the more soluble binding partner into the syringe at a high concentration.

Eq. 30.9 is valid for 1:1 binding of ligand to the macromolecule. For n independent binding sites, eq. 30.9 needs to be modified by multiplication of M_0 by n. The parameter n can be varied in the fitting procedure to determine the stoichiometry of the reaction. However, it is important to note that any error in the concentrations of the binding partners will affect n, and experimental values for n typically differ from integer numbers. If n is 0.9 or 1.1, for example, we may look at a 10% error in concentration determination for one binding partner while the concentration for the other is accurate. If both concentrations are inaccurate, n may not differ much from integer (if the errors compensate each other) or it may differ even more (if the errors add up). Accurate concentration determination of the binding partners is therefore essential for a reliable determination of binding stoichiometries by ITC.

The free energy of binding ΔG^0 can be calculated from K_a, and the change in entropy, ΔS, can be calculated according to the Gibbs-Helmholtz-equation from ΔG and ΔH (see Section 1.6, eq. 1.111). With these parameters, an almost complete set of thermodynamic parameters for the binding reaction is obtained from a single experiment. The last parameter missing in the thermodynamic profile is the heat capacity ΔC_p. The heat capacity is defined by the temperature dependence of ΔH (Section 1.4, eq. 1.62). ΔC_p can be obtained from a set of ITC experiments that determine ΔH at different temperatures (Figure 30.4). Due to the temperature dependence of ΔH, a single ΔH value measured at a certain temperature is of little value unless ΔC_p is known, or unless it is compared with other experiments that have been carried out at the same temperature.

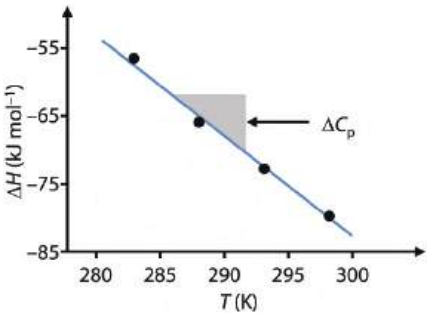


FIGURE 30.4: Heat capacity as the temperature dependence of ΔH. Data are for thymidine binding to thymidine kinase (www.bindingdb.org).

30.1.3 Origin of enthalpic changes

ΔH values for the binding of a ligand to a single site on a macromolecule are distributed over a wide range from +50 kJ mol^{-1} to –140 kJ mol^{-1}, although most values are in the range between –10 kJ mol^{-1} and –20 kJ mol^{-1} (data from www.bindingdb.org). The enthalpy measured in an ITC experiment is the sum of all heat changes that occur. Thus, ΔH is a composite of all enthalpic changes during each injection, including the breaking and formation of (hydrogen) bonds, van der Waals, ionic, and other polar and non-polar interactions (see Section 17.2.2). Decomposing the measured ΔH into individual molecular contributions is almost impossible. Apart from the binding event itself, protonation and/or deprotonation of the binding partners may occur. Heat changes associated with protonation events can be determined from a set of ITC experiments performed in different buffers at the same temperature and pH. If n protons are exchanged upon ligand binding, the total enthalpy ΔH is the enthalpy due to ligand binding ΔH_{ligand} plus n times the ionization enthalpy ΔH_{prot} of the buffer:

$$\Delta H = \Delta H_{ligand} + n \cdot \Delta H_{prot} \qquad \text{eq. 30.10}$$

To determine the number of protons that are exchanged, the observed enthalpy changes ΔH are plotted as a function of the tabulated buffer ionization enthalpy ΔH_{prot} (Table 30.1). Linear regression will give n as the slope and ΔH_{ligand} (the sum of all remaining contributions to ΔH) as the intercept with the y-axis.

TABLE 30.1:
ΔH and ΔC_p changes for the dissociation of protonated buffers in 0.1 M KCl at 25°C.

buffer	pK_A	ΔH_{prot} (kJ mol^{-1})	ΔC_p (J mol^{-1} K^{-1})
acetate	4.62	0.49 ± 0.02	2128 ± 2
MES	6.07	15.53 ± 0.03	16 ± 2
cacodylate	6.14	21.96 ± 0.02	278 ± 2
glycerol 2-phosphate	6.26	20.72 ± 0.02	2179 ± 2
PIPES	6.71	11.45 ± 0.04	19 ± 4
ACES	6.75	31.41 ± 0.05	227 ± 4
phosphate	6.81	5.12 ± 0.03	2187 ± 3
BES	7.06	25.17 ± 0.07	2 ± 5
MOPS	7.09	21.82 ± 0.03	39 ± 3
imidazole	7.09	36.59 ± 0.06	216 ± 5
TES	7.42	32.74 ± 0.03	233 ± 3
HEPES	7.45	21.01 ± 0.07	49 ± 5
EPPS	7.87	21.55 ± 0.05	56 ± 4
triethanolamine	7.88	33.59 ± 0.04	48 ± 3
tricine	8.00	31.97 ± 0.05	245 ± 4
bicine	8.22	27.05 ± 0.05	2 ± 4
TAPS	8.38	41.49 ± 0.06	23 ± 5
CAPS	10.39	48.54 ± 0.07	29 ± 6

Reproduced from Fukada & Takahasi (1998) Proteins 33(2):159–166 with permission.

Because many ligand binding reactions involve exchange of protons, the measured ΔH can be changed strongly by as much as 20 kJ mol^{-1} (compare PIPES and ACES in Table 30.1) simply by changing the identity of the buffer at the same pH. Sometimes, the overall (apparent) ΔH value can be further decomposed into contributions from metal binding or conformational changes of the protein.

In principle, the enthalpy change ΔH can also be determined from a series of ITC experiments at different temperatures. The van't Hoff equation (eq. 2.11) relates the temperature dependence of K_a to the enthalpy change ΔH, and ΔH can be determined from the slope of a van't Hoff plot (K_a *versus* $1/T$, see Figure 2.1). Van't Hoff plots are linear only if ΔC_p is constant over the respective temperature range (see Section 2.1.1). A difference between the van't Hoff enthalpy ΔH_{vH} and the enthalpy ΔH_{cal} determined directly from analysis of the calorimetric titration is an indication that other processes in addition to binding contribute to ΔH_{cal}. In the absence of such complications the two enthalpies should be identical.

30.1.4 Practical considerations

Since ITC directly measures heat, it is important that any change in heat that does not come from the binding event is corrected for. Common sources of unspecific heat generation are frictional heat, heats of dilution from unmatched buffers of the binding partners, and heats from side reactions, such as aggregation or unfolding.

Aggregates are a common source for frictional heat, and should be removed from all samples by centrifugation or at least by filtering prior to use. Aggregates can also form during the ITC experiment. Protein aggregation can often be prevented by a change in buffer composition, temperature, or pH. Some proteins aggregate when the solution is avidly stirred. In this case, the stirring speed of

the syringe can be reduced. To ensure that mixing is complete before the next injection, the time interval between injections might have to be increased. On the other hand, the resulting longer total experiment times may not be tolerated by aggregation-prone or unstable proteins. Sometimes, mild detergents or higher salt concentrations can help avoid aggregation by reducing hydrophobic and ionic interactions, respectively. The frictional heat from aggregates decreases the signal-to-noise ratio of the baseline, which makes them easy to detect (although the experiment has failed at this stage). Baseline correction is an essential step in data processing, and data that suffer from the presence of aggregates are difficult if not impossible to integrate. If protein stability continues to be a problem, the single-injection method might be a more suitable approach that allows the determination of thermodynamic parameters within minutes (Box 30.1).

BOX 30.1: SINGLE INJECTION METHOD.

A typical ITC experiment of 1 h duration can be shortened to about 20 min by the single injection method. Instead of using several discrete injections and waiting until the baseline is reached after each step, the binding partner is added in a single, slow, and continuous injection (Figure 30.5). This approach is only applicable if the association rate of the binding partners is sufficiently high, such that complex formation is complete within the 1–2 s response time of the instrument. The injection velocity in such an experiment is drastically decreased to 0.1 μL s^{-1} in order to ensure instant mixing of all reactant delivered by the syringe. The recorded curve is a binding isotherm in real time. The constant heat released at the beginning of the injection is caused by instantaneous binding of all ligand that is injected. At higher ligand concentrations, later in the injection, saturation is reached. The heat release decreases and finally returns to the baseline when the binding partner in the cell is saturated with ligand. For data analysis, the time scale is converted to ligand concentration, and the K_a value can be determined from the binding curve using eq. 30.9. The overall saving in experimental time needs to be balanced with the usually less accurate data from single injection experiments compared to the multiple injection method.

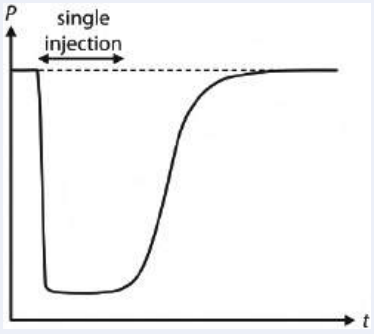


FIGURE 30.5: Single injection method. The duration of the single injection is indicated by a double arrow. During the injection, the power stays constant.

Toward the end of a multi-injection ITC experiment, when 100% complex has been formed, additional injections should lead to no or only very small heat changes. Any heat change that is observed after saturation has been reached can typically be attributed to heats of dilution. Sizeable heats of dilution arise when the buffers of the binding partners do not match exactly with respect to concentration, ionic strength, or pH, and such differences should hence be avoided. To ensure identical buffer conditions for both binding partners, macromolecular binding partners should be dialyzed into the same buffer. Small molecule ligands that cannot be dialyzed should be as pure as possible and be dissolved in the same buffer as the macromolecule. It is important to monitor

the pH during dilution of small molecules. Concentrated nucleotide solutions, for example, can easily change the pH of buffers with low capacity. If one of the two binding partners is not very soluble, it should be placed in the cell where a lower concentration is sufficient. Some small molecule ligands are only soluble in organic solvents, often DMSO, at the concentrations required for ITC. Organic solvents, particularly DMSO, have large heats of dilution. In this case, matching the buffers by including similar concentrations of organic solvents in the syringe and sample cell is necessary. If buffer matching is not possible, the heats of dilutions can be estimated in separate titrations of the binding partners into buffer and subtracting the two reference datasets from the binding data.

Side-reactions can also affect the measured heat. Unfolding of a protein during the ITC experiment adds the ΔH of unfolding to the measured heat and severely distorts the signal. Although ITC cells are made of inert materials, unstable or hydrophobic proteins may unfold and attach to the walls of the cell. This effect can be exacerbated by the presence of high detergent concentrations that are necessary to stabilize membrane proteins in aqueous buffers. Unfolding usually leads to a smooth baseline drift. As unfolding changes the total concentration of the active molecule in the cell, the determined values for ΔH, K_a, and the stoichiometry n are incorrect. Chemical reactions in the sample cell also pose problems. For example, a contaminating protease may generate heat by catalyzing hydrolysis of the binding partners.

In principle, equilibrium dissociation constants from the millimolar to the nanomolar range can be measured by ITC, but the accuracy of these constants depends on the concentrations of the binding partners used in the experiments. As a rule of thumb, the binding partner in the cell should be 10-fold more concentrated than the expected K_d, and the concentration of the molecule in the syringe should be 10-fold larger than that of the molecule in the cell. Under these conditions, the volume change in the cell due to repeated injections is kept small, and the calculation of ΔH, K_a, and n from the shape of the binding curve is straightforward. A more quantitative measure to design an ITC experiment is the Wiseman constant C, the product of the association constant K_a and the concentration of the molecule in the cell M_0. The Wiseman constant influences the shape of the binding curve and hence its information content (Figure 30.6).

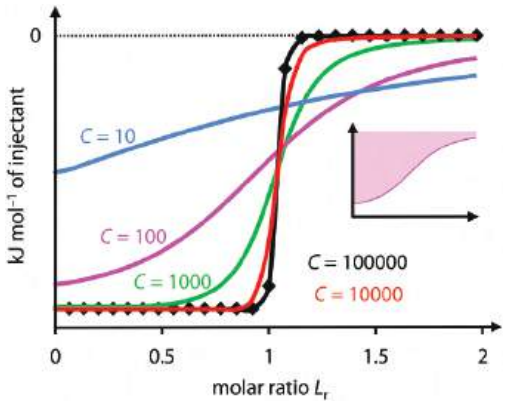


FIGURE 30.6: Wiseman constant and the shape of ITC curves. Simulated ITC data for 1:1 stoichiometry, M_0 = 0.1 mM, and K_a varying by a factor of ten from 10^9 M^{-1} (black curve with diamonds; C = 10 000) to 10^5 M^{-1} (flat, blue curve; C = 10). The high-affinity measurement has no data points in the transition region. The flat low-affinity data does not allow accurate estimation of n, or K_a. The green curve with C = 1000 represents the best combination of K_a and M_0. From the curves with C = 10 or C = 100, ΔH cannot be obtained from the y-axis intercept, but has to be calculated from the total area under the curve (inset).

When the Wiseman constant C is between 100 and 1000, binding curves show a pronounced sigmoidal shape and are moderately steep. Binding curves that are too steep will not have sufficient data points (if any) in the transition region, making determination of K_a difficult. Curves that are too flat, on the other hand, will result in large errors on n, ΔH, and K_a. For a binding curve with a Wiseman constant of C = 100, the enthalpy change ΔH cannot be derived from the intercept of the binding curve with the y-axis. In these cases, ΔH can still be obtained from the total area under the binding curve (Figure 30.6). K_a and n, on the other hand, cannot be determined accurately.

30.1.5 Measuring high affinities with ITC by competition

The higher the affinity (the larger K_a), the lower is the required concentration of the binding partner in the sample cell to reach a Wiseman constant between 100 and 1000. With decreasing concentrations, the heat exchanged becomes smaller and smaller, and the measured signal might become too small to be detected reliably. In this case, displacement of a weakly bound ligand by a high-affinity ligand is an alternative to still determine the K_a value. For this type of experiment, a low-affinity ligand B with known $K_{a,low}$ and ΔH_{low} is added to the macromolecule in the measurement cell, and the resulting low-affinity complex is then titrated with the high-affinity ligand. The heat is calculated from the integrated peaks to generate the displacement titration curve that can be analyzed for the true binding constant of the high-affinity ligand. The known $K_{a,low}$ value for the low-affinity ligand is fixed during data analysis. The concentration of the high-affinity complex is described by the solution of a cubic equation (see Section 21.5.5.1, Box 21.10). The apparent high-affinity binding constant K_{app} is given by

$$K_{app} = \frac{K_{a,high}}{1 + K_{a,low} \cdot [B]} \qquad \text{eq. 30.11}$$

The apparent binding constant K_{app} is smaller than $K_{a,high}$ for the high-affinity interaction and can therefore be measured with greater accuracy. By adjusting the concentration of the low-affinity ligand $[B]$ in the sample cell, the apparent binding constant K_{app} can be modulated. The apparent enthalpy change is approximated by

$$\Delta H_{app} = \Delta H_{high} - \Delta H_{low} \cdot \frac{K_{a,low} \cdot [B]}{1 + K_{a,low} \cdot [B]} \qquad \text{eq. 30.12}$$

If the low-affinity ligand B is present in large excess over its binding partner, the free concentration $[B]$ in eq. 30.11 and eq. 30.12 can be replaced by the known total concentration of B_0.

30.1.6 Measuring Michaelis-Menten enzyme kinetics with ITC

ITC can also be used to measure the steady-state enzyme kinetic parameters k_{cat} and K_M for enzymatic reactions (Sections 13.1; 13.2). The enzymatic reaction must be associated with a heat change and must be slow on the time scale of the instrument response. Typically, the enzyme is placed in the measurement cell, and the substrate is injected from the syringe. The heat is released over a larger time window compared to a titration experiment, while the enzyme turns over its substrate. Once all substrate has been converted to products, the signal returns to the baseline (Figure 30.7).

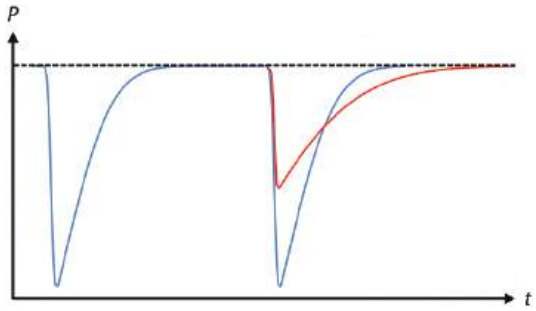


FIGURE 30.7: Enzyme reactions in ITC. Schematic raw ITC data upon injection of a small amount of substrate into a reaction cell containing an enzyme. The peak areas are larger than for a single injection in a binding equilibrium, and the signal takes much longer to return to the baseline. A smaller signal in a second, identical injection indicates whether inhibition of the enzyme by the product generated during the first injection occurs. The red curve would indicate such product inhibition.

The experiment is useful to derive ΔH of the reaction and to test for product inhibition by a second injection of substrate. If the second heat change is identical to the first, no product inhibition has occurred. This aspect is further discussed later in this Section. To evaluate the data we differentiate eq. 30.8 by time (product formation now substitutes for the complex ML) and see how the feedback power depends on the formation of product P in the measuring cell:

$$power = \frac{dq}{dt} = \frac{d[P]}{dt} \cdot \Delta H \cdot V_0 \qquad \text{eq. 30.13}$$

Integration of the power signal dq/dt over the time the system needs to return to the baseline yields the total heat ΔH evolved during the enzymatic reaction:

$$\Delta H = \frac{1}{V_0 \cdot [P]} \cdot \int_{t=0}^{\infty} \frac{dq}{dt} dt \qquad \text{eq. 30.14}$$

The determination of ΔH is possible even if the substrate concentrations are not saturating. In order to obtain Michaelis-Menten parameters, the substrate concentration in the reaction cell has to be increased such that the enzyme is saturated. Subsequent injections are performed well before all substrate has been turned over. Upon each further injection, the baseline is then shifted in a stepwise manner (Figure 30.8).

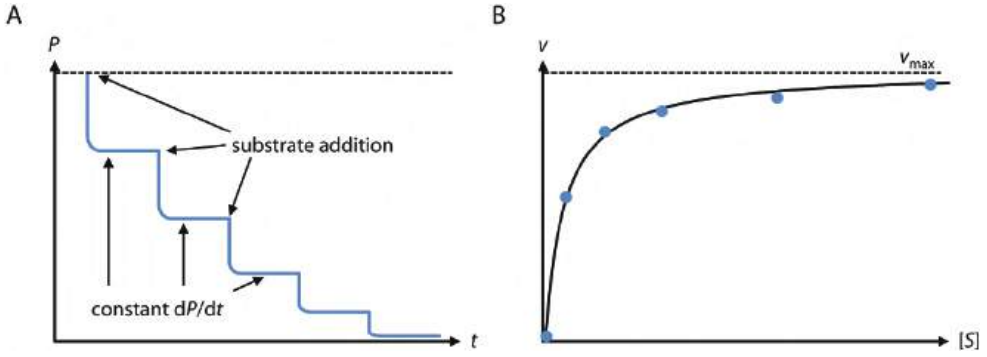


FIGURE 30.8: Schematic ITC data upon injection of an excess of ligand into a reaction cell containing an enzyme. A: The continuous reaction under steady-state conditions shifts the baseline after each substrate injection. B: The rates calculated from the baseline shifts are plotted as a function of substrate concentration [*S*]. Note that ΔH cannot be derived from this type of experiment because the substrate is not completely converted prior to starting the next injection. ΔH needs to be estimated from an experiment that reaches the baseline (Figure 30.7).

The shifts of the baseline can be converted into reaction velocities according to eq. 30.15, which we obtain by rearrangement of eq. 30.13 and equating to the Michaelis-Menten equation. Under substrate saturation the reaction velocity v is constant (Figure 30.8).

$$v = \frac{d[P]}{dt} = \frac{1}{\Delta H \cdot V_0} \frac{dq}{dt} = \frac{k_{cat} \cdot E_0 \cdot [S]}{[S] + K_M} \qquad \text{eq. 30.15}$$

The reaction velocities are plotted as a function of substrate concentration [*S*], and the curve obtained is described by eq. 30.15 to determine k_{cat} and K_M. For such experiments, the data need to be corrected for dilution of the total enzyme concentration E_0 by the displacement of solution from the cell in every injection.

A continuous enzyme assay is also possible with ITC. In this type of experiment, a large amount ($[S] \gg K_M$) of substrate is injected, such that product is formed at the maximum rate v_{max}. The signal is recorded until substrate depletion, where the power signal returns to the baseline (Figure 30.9).

Under these conditions, inhibition of the enzyme can be investigated by performing the reaction at different concentrations of inhibitor. The areas under curves with or without inhibitor are identical, yielding ΔH of the reaction (eq. 30.14). In addition, the inhibition constant K_i can be calculated from the reaction velocities (see Section 13.6).

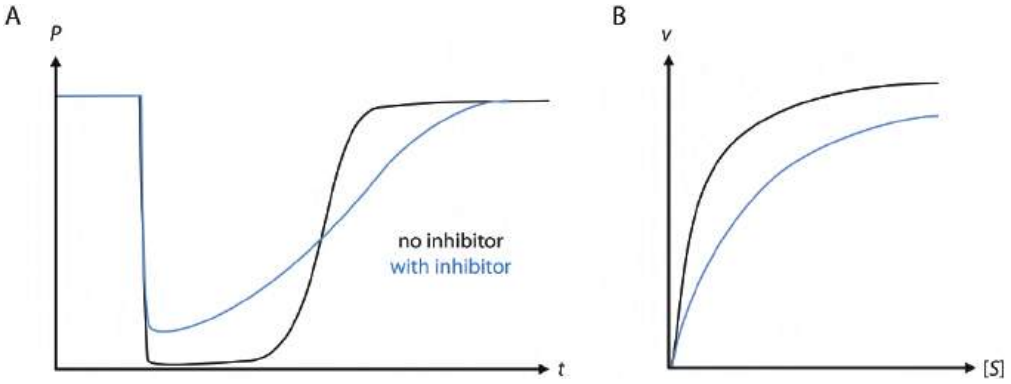


FIGURE 30.9: Substrate turnover and enzyme inhibition. A: Power signal for substrate turnover by an enzyme without (black) and with (blue) inhibitor. B: Conversion of the heat signal to reaction velocities as a function of substrate concentration [*S*].

The actual substrate concentration $[S](t)$ at any time t along the curve can be calculated by subtraction of the actual product concentration $[P](t)$ from the total substrate concentration S_0 that was injected. $P(t)$ can be calculated from eq. 30.14, which gives

$$[S](t) = S_0 - [P](t) = S_0 - \frac{\int_{t=0}^{t} \frac{dq}{dt} dt}{\Delta H \cdot V_0} \qquad \text{eq. 30.16}$$

The resulting hyperbolic curves can be described by the Michaelis-Menten equation (right-hand side of eq. 30.15) to obtain k_{cat} and K_M in the absence and presence of inhibitor.

Kinetic information on binding and dissociation reactions can also be obtained from ITC experiments by analysis of the shape of the individual injection profiles (Box 30.2).

BOX 30.2: PRE-STEADY-STATE KINETIC INFORMATION FROM ITC – KINITC.

The analysis of ITC curves for thermodynamic parameters described above only uses the information of the integrated heat, but does not take into account the shape of the curve, *i.e.* the velocity with which the signal returns to the baseline. The shape of this signal is determined by the instrument response, by the finite duration of injection and mixing, and by the rate constants of the reaction. If the response time of the instrument as well as the duration of the injection and mixing are shorter than the time scale of the association reaction, the rate constants of the binding equilibrium can be determined by analyzing the return of the signal to the baseline. The fixed instrument response therefore puts an upper limit on the rate constants that can be determined. To extract the rate constants for binding and dissociation, a global analysis of the return to the baseline from all peaks can be performed in terms of a kinetic 1:1 binding model. This approach is termed kinetic ITC or kinITC, a label-free method that has the advantage of not requiring a spectroscopic signal (Figure 30.10). Burnouf et al. (2012). J. Am. Chem. Soc. 134(1): 559–565.

(Continued)

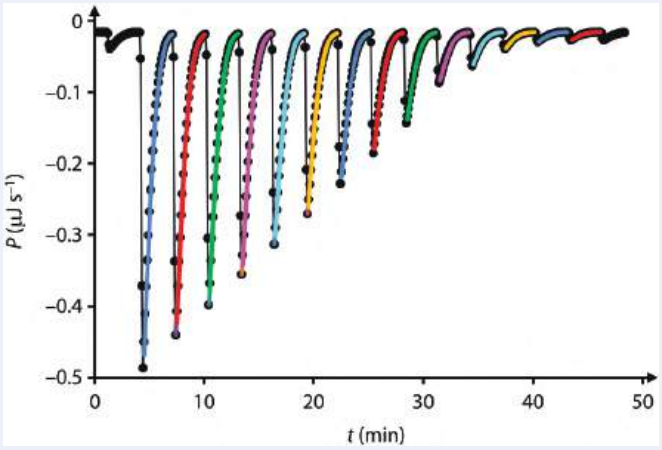


FIGURE 30.10: kinITC. Only the exponential parts of the power signal are evaluated by exponential fits.

30.2 DIFFERENTIAL SCANNING CALORIMETRY

We have seen in Section 30.1 how ITC is used to measure the heat associated with binding or catalytic events. In contrast, *differential scanning calorimetry* (DSC) measures the heat released or taken up during structural or phase transitions. DSC is most often used to study temperature-induced protein and nucleic acid folding and unfolding. DSC directly measures thermodynamic quantities of structural transitions, and does not rely on a spectroscopic probe. Measurements can be performed under pressure, which expands the experimentally accessible temperature range to > 100°C.

30.2.1 General principle

Similar to ITC instruments, a DSC instrument consists of a reference cell, filled with buffer, and a sample cell, filled with the molecule of interest in buffer, and both cells are enclosed by an adiabatic jacket. In contrast to ITC, DSC is not isothermal but both cells are heated during the experiment and held at a constant temperature difference ΔT by an electrical feedback system. When a structural transition such as a protein unfolding event takes place in the sample cell, heat is absorbed (protein unfolding is endothermic). To maintain a constant temperature difference ΔT between the reference and sample cells, the heating power to the sample cell is adjusted; in this case it is increased. The heating power is the primary signal and is proportional to the overall (excess) heat capacity C_p, which represents the temperature dependence of the enthalpy H. The raw data of a DSC measurement is a curve of C_p as a function of T, the *thermogram*. After correction for the instrument baseline (caused by the heat capacity of the instrument itself), this thermogram displays a linear baseline before and after the structural transition, and a bell-shaped peak for the transition (Figure 30.11). The instrument baseline is recorded in a separate experiment with buffer in the sample cell and subtracted from the raw DSC data.

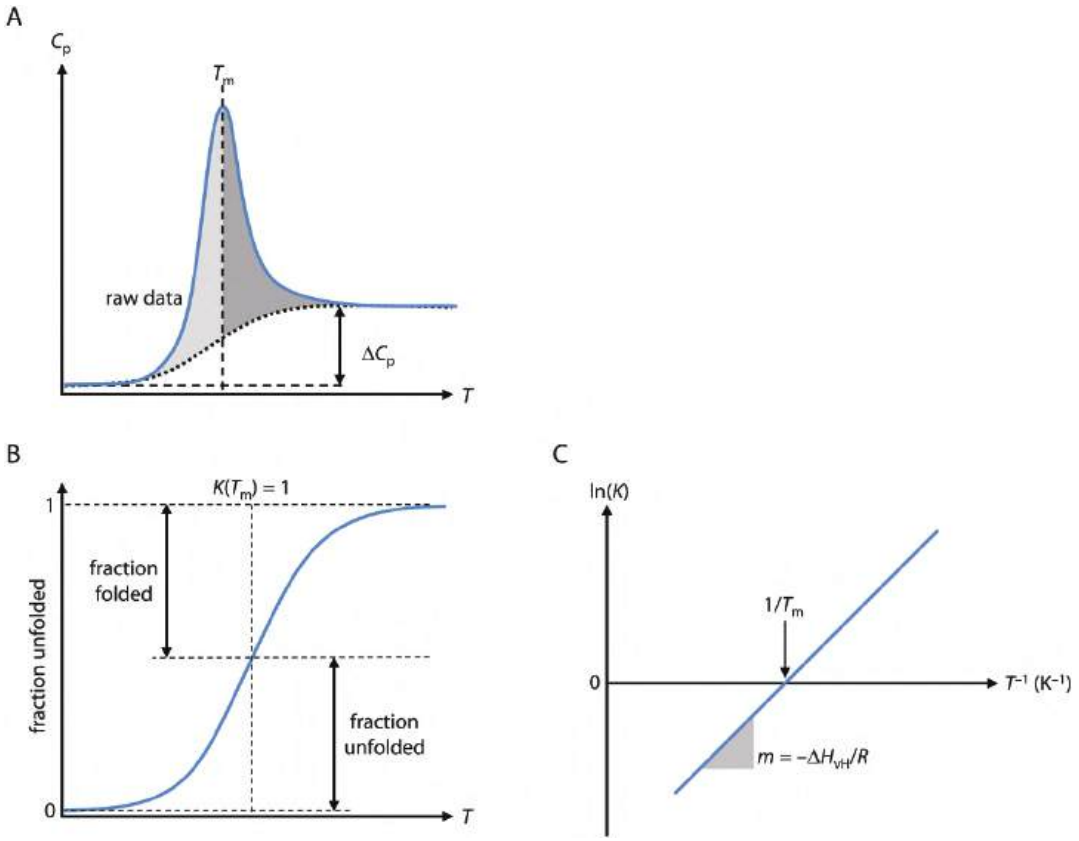


FIGURE 30.11: Schematic DSC thermograms. A: Endothermic two-state protein unfolding transition corrected for the instrument baseline. The total area under the curve yields the enthalpy change ΔH of the reaction. If the offsets of the two baselines before and after the experiment are different, there is an overall change in heat capacity (ΔC_p). Should the slopes of the pre- or post-transition baselines not be zero, then ΔC_p itself is temperature-dependent. The areas shaded in gray correspond to the fractions of folded and unfolded protein at a given temperature. B: At any given temperature, the ratio of the areas in (A) is the equilibrium constant. Plotting the unfolded fractions as a function of temperature results in an unfolding curve. C: The van't Hoff plot resulting from the equilibrium constants $K(T)$ in panel (B). The enthalpies from the thermogram and the van't Hoff plot should be identical.

T_m, the "melting" temperature of the molecule studied, is accurately obtained from the peak of the thermogram. For single-domain proteins, T_m is the temperature at which half of the molecules are unfolded. T_m correlates with the overall protein stability. It should be similar to T_m values from temperature-induced unfolding experiments followed by spectroscopic methods such as CD, UV, and fluorescence. Further analysis of thermograms to obtain thermodynamic parameters is only valid if the measured transitions are reversible. To test for reversibility, a second thermogram is measured after the first, where the temperature is gradually decreased from the highest to the lowest temperature. For reversible transitions both thermograms are identical, whereas differences are indicative of irreversible transitions. In this case, the information content of the thermogram is limited to the apparent melting temperature T_m of the first transition.

The total change in heat capacity ΔC_p during the transition is obtained from the difference between the baselines. A positive value for ΔC_p (Figure 30.11) indicates solvation of non-polar surfaces and is a hallmark for protein unfolding. Indeed, ΔC_p seems to be the thermodynamic parameter dominating the hydrophobic effect, which drives protein folding. For some proteins, a correlation between ΔC_p and the amount of (non-polar) surface area that is buried upon complex formation or protein folding has been inferred, but overall it is difficult to relate surface properties to heat capacity. On the other hand, negative values for ΔC_p signify solvation of polar surfaces, typically seen with sequence-specific binding of proteins to DNA. For example, binding of the *trp* repressor

to the *trp* operator DNA has a large negative ΔC_p of almost −4 kJ mol^{-1} K^{-1}. By contrast, non-specific DNA binding by proteins does not change ΔC_p very much.

For reversible transitions, the thermodynamic parameters for the structural transition can be determined from the thermogram. The change in heat capacity ΔC_p during the transition is obtained from the difference between the baselines before and after the transition. The overall change in enthalpy ΔH is obtained from the peak area by integration (Figure 30.11) and normalization for protein concentration.

$$\Delta H(T) = \int_{T_0}^{T} C_p(T)\,dT \qquad \text{eq. 30.17}$$

Because the baseline after the transition often does not return to the value before the transition, the shape of the baseline during the transition needs to be extrapolated before the area under the peak can be integrated to determine ΔH. The most basic model assumes a jump of the baseline from the value before the transition to the final value at the melting temperature T_m, whereas more sophisticated models involve a linear or quadratic interpolation. However, for every data point in the transition region, a mixture of two states is present, and the baseline is perhaps best assumed as a sigmoidal curve (Figure 30.11) that reflect the progress of the protein unfolding transition (see Figure 18.57). The choice of baseline model affects the area of the transition, and hence the value of ΔH, but these errors are often smaller than other uncertainties such as inaccurate determination of protein concentrations.

The equilibrium constant $K(T)$ for the structural transition can be calculated for any given temperature in the experiment within the transition region (Figure 30.11). For each temperature T, the two peak areas below and above T are integrated separately. The area above T (right-hand side) reflects the fraction of folded protein, the area below T (left-hand side) represents the fraction that has unfolded at this temperature. The ratio of these areas corresponds to $K(T)$. As usual, $\Delta G(T)$ can be calculated from $K(T)$, and the extrapolation to the standard temperature yields ΔG^0. Similar to ITC, ΔS^0 is then calculated according to the Gibbs-Helmholtz equation from ΔG^0 and ΔH^0. The complete set of thermodynamic parameters for the structural transition is obtained from one experiment, this time including ΔC_p.

Sometimes multiphasic transitions are observed (Figure 30.12). In this case, the thermogram needs to be described by two or more peaks, and individual ΔC_p, T_m, and ΔH values, as well as $K(T)$, $G(T)$, and $S(T)$ for each transition are determined.

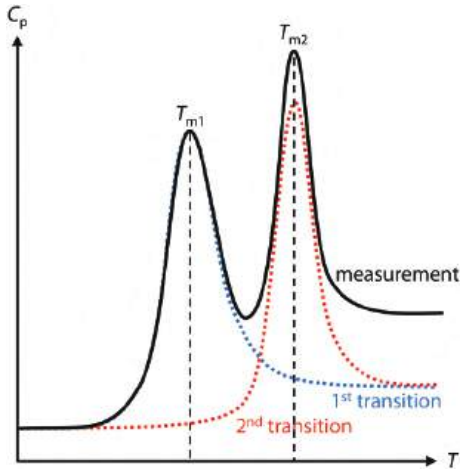


FIGURE 30.12: Multiple transitions. The separate unfolding of two domains is shown. When the T_m values are very different, the sequential unfolding of a two-domain protein can be deconvoluted into single transitions (blue and red dashes).

30.2.2 Two-state unfolding of macromolecules

DSC can measure the cooperative unfolding of any macromolecule. We would now like to derive a fitting function to describe a DSC thermogram. In the following, we consider a two-state protein unfolding of the form:



scheme 30.1

N and U are the folded and unfolded states, respectively. The equilibrium constant for unfolding can be expressed in terms of either the concentrations of the folded and unfolded states or the fraction of unfolded protein f_U:

$$K_{\text{unfold}} = K_U = \frac{[U]}{[N]} = \frac{f_U}{1 - f_U} \qquad \text{eq. 30.18}$$

Solving eq. 30.18 for f_U yields

$$f_U = \frac{K_U}{1 + K_U} \qquad \text{eq. 30.19}$$

The fraction of unfolded protein helps us in the next step to describe the total enthalpy of the system. The Hess law states that the enthalpies of a system are additive (Box 1.5). Thus, the total molar enthalpy H in the DSC cell is

$$H(T) = H_N(T) + f_U \cdot \Delta H_U(T) \qquad \text{eq. 30.20}$$

The total enthalpy is the enthalpy of the folded state, $H_N(T)$, plus the enthalpy change of unfolding, $\Delta H_U(T)$, weighted by the fraction of unfolded protein f_U. If there are several unfolding transitions, an equivalent number of terms $f_{U,i} \cdot \Delta H_i(T)$ has to be added to eq. 30.20. The total molar heat capacity at constant pressure, C_p, is the temperature derivative of eq. 30.20:

$$C_p(T) = C_{pN} + f_U \cdot \Delta C_{pU}(T) + \Delta H_U(T) \cdot \left(\frac{\partial f_U}{\partial T} \right) \qquad \text{eq. 30.21}$$

$\Delta C_{pN}(T)$ and $\Delta C_{pU}(T)$ are the heat capacities of the folded and unfolded states, respectively. In this equation, the differential $\partial f_U / \partial T$ can be calculated by differentiating eq. 30.19:

$$\frac{\partial f_U}{\partial T} = \frac{1}{(1 + K_U)^2} \cdot \frac{\partial K_U}{\partial T} \qquad \text{eq. 30.22}$$

Realizing that K is an exponential function of the temperature, we can re-write the differential $\partial K_U / \partial T$:

$$\frac{\partial K_U}{\partial T} = K_U \cdot \frac{\partial \ln K_U}{\partial T} \qquad \text{eq. 30.23}$$

The van't Hoff enthalpy is given by the temperature dependence of the equilibrium constant:

$$\frac{\partial \ln K_U}{\partial T} = \frac{\Delta H_{vH}}{RT^2} \qquad \text{eq. 30.24}$$

So we can re-write eq. 30.21

$$C_p(T) = C_{pN}(T) + \frac{K_U}{1 + K_U} \cdot \Delta C_{pU}(T) + \Delta H_U(T) \cdot \frac{K_U}{(1 + K_U)^2} \cdot \frac{\Delta H_{vH}}{RT^2} \qquad \text{eq. 30.25}$$

This equation is the general formulation of the heat capacity of a two-state system at a specific temperature. More transitions require extra terms for each K_U, so eq. 30.25 can be adapted for more complicated unfolding scenarios. If we assume that the van't Hoff enthalpy equals the calorimetric enthalpy of unfolding $\Delta H_U(T)$, eq. 30.25 simplifies to

$$C_p(T) = C_{pN}(T) + \frac{K_U(T)\cdot\Delta C_{pU}(T)}{1+K_U(T)} + \frac{K_U\cdot\left[\Delta H_U(T)\right]^2}{\left(1+K_U(T)\right)^2\cdot RT^2} \qquad \text{eq. 30.26}$$

All that we have to do now is to express $\Delta H_U(T)$ and $K_U(T)$ in terms of the heat capacity and some easily accessible parameters. We choose as parameters the enthalpy of unfolding $\Delta H_U(T_m)$ at the transition midpoint temperature T_m (Figure 30.11). We have derived expressions for $\Delta H_U(T)$ and $K_U(T)$ before in this book. $\Delta H_U(T)$ is accessible from integration of Kirchhoff's law (eq. 1.69):

$$\int_{T_m}^{T} d(\Delta H_U) = \int_{T_m}^{T} \Delta C_{pU} dT \qquad \text{eq. 30.27}$$

This yields

$$\Delta H_U(T) = \Delta H_U(T_m) + \Delta C_{pU}\cdot(T - T_m) \qquad \text{eq. 30.28}$$

$\Delta H_U(T_m) = \Delta H_m$ is the associated enthalpy of unfolding at the melting temperature T_m. From eq. 18.29 (see Section 18.5.4), we know an expression for $\ln K_U(T)$:

$$-RT\ln K_U(T) = \Delta H_m\cdot\left(1-\frac{T}{T_m}\right) + \Delta C_{pU}\cdot\left(T - T_m - T\cdot\ln\frac{T}{T_m}\right) \qquad \text{eq. 30.29}$$

$K_U(T)$ is readily obtained from eq. 30.29 by rearrangement:

$$K_U = e^{-\frac{\Delta H_m}{RT}\left(1-\frac{T}{T_m}\right)-\frac{\Delta C_{pU}}{RT}\left(T-T_m-T\cdot\ln\frac{T}{T_m}\right)} \qquad \text{eq. 30.30}$$

Substituting eq. 30.29 and eq. 30.30 into eq. 30.26 and expressing the temperature dependence of $\Delta C_{pN}(T)$ by a linear function, the baseline, as $\Delta C^0_{pN} + m_N\cdot T$, yields quite a complex function that we will refrain from reproducing here. It allows fitting of the measured DSC data with the five fit parameters ΔC^0_{pN}, m_N, T_m, ΔH_m, and ΔC_{pU}. If necessary, a second baseline for the unfolded state can be added to the fitting equation. Baseline correction can also be done before fitting, so the values of $C_{pN}(T)$ and ΔC_{pU} are effectively set to zero, and the $C_p(T)$ equation (eq. 30.25) simplifies to

$$C_p(T) = \frac{K_U\cdot\left[\Delta H_U(T)\right]^2}{\left(1+K_U(T)\right)^2\cdot RT^2} \qquad \text{eq. 30.31}$$

with

$$\Delta H_U(T) = \Delta H_U(T_m) = \Delta H_m \qquad \text{eq. 30.32}$$

and

$$K_U = e^{-\frac{\Delta H_m}{RT}\left(1-\frac{T}{T_m}\right)} \qquad \text{eq. 30.33}$$

Substituting eq. 30.32 and eq. 30.33 into eq. 30.31, we arrive at the more manageable expression for $C_p(T)$ that has only T_m and ΔH_m as fit parameters:

$$C_p(T) = \frac{e^{-\frac{\Delta H_m}{RT}\left(1-\frac{T}{T_m}\right)} \cdot \left[\Delta H_m\right]^2}{\left(1+e^{-\frac{\Delta H_m}{RT}\left(1-\frac{T}{T_m}\right)}\right)^2 \cdot RT^2} \qquad \text{eq. 30.34}$$

30.2.3 Two-state unfolding with subunit dissociation

A frequently observed transition that is amenable to DSC is the coupled dissociation of oligomers and unfolding of the constituent monomers. In contrast to the separate transitions that are possible with multi-domain proteins (Figure 30.12), the DSC thermogram in this case shows only a single transition. A slight variation of the expression of $K(T)$ accommodates the oligomer-monomer equilibrium. We assume an oligomer of n subunits N_n that simultaneously dissociates and unfolds into n unfolded monomers U.

$$N_n \overset{K_{unfold}}{\rightleftharpoons} n\,U \qquad \text{scheme 30.2}$$

The overall equilibrium constant is then

$$K = \frac{[U]^n}{[N_n]} = \frac{f_U}{1-f_U} \qquad \text{eq. 30.35}$$

The total molar concentration in terms of oligomers is

$$c_{tot} = [N_n] + \frac{[U]}{n} \qquad \text{eq. 30.36}$$

and the fraction f_U of unfolded monomer is:

$$f_U = \frac{[U]}{n \cdot c_{tot}} \qquad \text{eq. 30.37}$$

Combining eq. 30.36 and eq. 30.37 to express $[N_n]$ and $[U]$ in terms of f_U yields the following expression for K:

$$K = \frac{{f_U}^n}{1-f_U} \cdot n^n \cdot {c_{tot}}^{n-1} \qquad \text{eq. 30.38}$$

We know that at the melting temperature T_m, the fraction of unfolded monomers is 0.5:

$$K(T_m) = 0.5^{n-1} \cdot n^n \cdot {c_{tot}}^{n-1} \qquad \text{eq. 30.39}$$

Unlike in the previous two-state unfolding transition (eq. 30.29), $K(T_m)$ is now different from unity. In a dimer-monomer dissociation/unfolding equilibrium with $n = 2$, $K(T_m)$ equals $2 \cdot c_{tot}$. Thus, $K(T)$ from eq. 30.29 modifies to

$$K(T) = 0.5^{n-1} \cdot n^n \cdot {c_{tot}}^{n-1} \cdot e^{-\frac{\Delta H_m}{RT}\left(1-\frac{T}{T_m}\right)-\frac{\Delta C_{pU}}{RT}\left(T-T_m-T\cdot\ln\frac{T}{T_m}\right)} \qquad \text{eq. 30.40}$$

ΔC_{pU} in the equation above represents the change in heat capacity that accompanies both dissociation of the oligomer and coupled unfolding of the monomers. The same holds true for ΔH_m.

Once $K(T)$ is known from eq. 30.40, numerical evaluation is required to solve eq. 30.38 for $f_U(T)$. The temperature-dependent change $\partial f_U/\partial T$ in eq. 30.21 is then calculated and used to describe $C_p(T)$ of the DSC experiment. The same simplifications with respect to baseline subtraction and temperature-independence of ΔC_{pU} as discussed above can be applied to this case.

It should be noted that in addition to the two-state model of protein unfolding, alternative models to analyze DSC thermograms using statistical thermodynamics have been developed. These models do not assume temperature-independence of ΔH and ΔS, and make no assumptions on the shape of the baseline. Numerical integration of the thermograms yields the partition function of the system, from which in addition to the heat capacities of the native and unfolded states the temperature dependence of ΔH and ΔS can be extracted.

QUESTIONS

30.1 An ITC with a cell volume of 1.4 mL is filled with 10 µM protein solution. The injection volumes are 5 µL of a 50 µM ligand solution every 2 min. After the titration is finished and evaluated, a K_a of 10^9 M^{-1} and ΔH of –40 kJ mol^{-1} is found. What is the heat evolved during the first injection and what average power is recorded by the instrument?

30.2 For the competitive titration of a low-affinity protein-ligand complex MB characterized by $K_{a,low}$ with a high-affinity ligand A, the apparent equilibrium constant is

$$K_{app} = \frac{K_{a,high}}{1 + K_{a,low} \cdot [B]} \tag{1}$$

Using the laws of mass action for the two equilibria, derive the formula above.

30.3 The data given in the table below are from thymidine binding to thymidine kinase. Estimate the heat capacity from these data.

T (°C)	ΔH (kJ mol^{-1})
25	–79.91
20	–73.08
15	–66.36
10	–57.12

30.4 ITC of a protein-ligand interaction at 298 K, pH 7, and 150 mM KCl using buffers of different deprotonation enthalpies ΔH_{prot} for the BH^+ ion yields the following enthalpies $\Delta H_{measured}$:

buffer	pK_A	ΔH_{prot} (kJ mol^{-1})	$\Delta H_{measured}$ (kJ mol^{-1})
PIPES	6.71	11.4	–10.1
ACES	6.75	31.4	–29.4
Phosphate	6.81	5.1	–3.4
MOPS	7.09	21.8	–20.2
imidazole	7.09	36.6	–34.4

Are these buffers suitable for this measurement? Is there proton transfer during the binding event and, if so, how many protons are exchanged and in which direction?

30.5 Micelle formation follows to a good approximation an all-or-nothing equilibrium:

$$\mathrm{n \cdot S} \rightleftarrows \mathrm{M_n}$$

with the equilibrium constant K given as

$$K = \frac{[M_\mathrm{n}]}{[S]^n}$$

which leads to the relation

$$[S]_\mathrm{tot} = [S] + n \cdot K \cdot [S]^n$$

The change in amphiphile concentration over the total concentration that is measured in ITC is given by the first derivative of the equation above:

$$\frac{d[S]}{d[S]_\mathrm{tot}} = \frac{[S]}{[S] + n^2 \cdot K \cdot [S]^n} = \frac{1}{1 + n^2 \cdot K \cdot [S]^{n-1}}$$

Show that at the equivalence concentration, where the ITC experiment usually stops and the condition $[S] = n \cdot [M_\mathrm{n}]$ applies, the aggregation number n of the micelle may be calculated from the slope of the curve as

$$\frac{d[S]}{d[S]_\mathrm{tot,equiv}} = \frac{1}{1+n}$$

REFERENCES

Armstrong, K. M. and Baker, B. M. (2007) A comprehensive calorimetric investigation of an entropically driven T cell receptor-peptide/major histocompatibility complex interaction. *Biophys J* **93**(2): 597–609.
· binding of a T-cell receptor to a peptide-MHC complex is analyzed by ITC – the enthalpy is a composite of metal binding and/or conformational changes

Burnouf, D., Ennifar, E., Guedich, S., Puffer, B., Hoffmann, G., …, Dumas, P. (2012) kinITC: A new method for obtaining joint thermodynamic and kinetic data by isothermal titration calorimetry. *J Am Chem Soc* **134**(1): 559–565.
· introduces the idea of obtaining kinetic information from ITC data

Fukada, H. and Takahashi, K. (1998) Enthalpy and heat capacity changes for the proton dissociation of various buffer components in 0.1 M potassium chloride. *Proteins* **33**(2): 159–166.
· treatment of changes in enthalpy and heat capacity for common buffers

Horn, J. R., Russell, D., Lewis, E. A. and Murphy, K. P. (2001) Van't Hoff and calorimetric enthalpies from isothermal titration calorimetry: Are there significant discrepancies? *Biochemistry* **40**(6): 1774–1778.
· shows that the van't Hoff enthalpy and ΔH from ITC should not differ

Ladbury, J. E., Wright, J. G., Sturtevant, J. M. and Sigler, P. B. (1994) A thermodynamic study of the *trp* repressor-operator interaction. *J Mol Biol* **238**(5): 669–681.
· binding of the *trp* repressor to its operator DNA sequence has a large negative heat capacity

Prabhu, N. V. and Sharp, K. A. (2005) Heat capacity in proteins. *Annu Rev Phys Chem* **56**: 521–548.
· review on heat capacity with special attention to protein folding

Seelig, J. and Seelig, A. (2023) Protein unfolding – Thermodynamic perspectives and unfolding models. *Int J Mol Sci* **24**: 5457–5474.
· review on alternative models for analyzing DSC thermograms

Velazquez-Campoy, A., Ohtaka, H., Nezami, A., Muzammil, S. and Freire, E. (2004) Isothermal titration calorimetry. *Curr Protoc Cell Biol* **17**: Unit 17.8.
· practical review on ITC

Wiseman, T., Williston, S., Brandts, J. F. and Lin, L. N. (1989) Rapid measurement of binding constants and heats of binding using a new titration calorimeter. *Anal Biochem* **179**(1): 131–137.
· introduction of the first microcalorimeter

Appendix

31

Mathematical concepts used in this book

The mathematics included in this chapter is treated only to the depth necessary to follow the transformations described in the book and to solve the questions. The chapter is incomplete, and many of the topics will be familiar to the reader. They are merely summarized here for quick reference.

31.1 SUMS AND PRODUCTS

Infinite sums and products can be represented in short forms:

sum:
$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n \qquad \text{eq. 31.1}$$

For example, the root mean square distance between two superimposed molecules is defined as the following sum:

$$rmsd = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \delta_i^2} \qquad \text{eq. 31.2}$$

where δ_i is the distance between the i^{th} atom pair and N is the total number of atom pairs chosen for the calculation, *e.g.* equivalent Cα atoms or backbone atoms in proteins and phosphorous or ribose atoms in nucleic acids.

product:
$$\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdot \ldots \cdot x_n \qquad \text{eq. 31.3}$$

A special product is the factorial, which is defined as

$$\prod_{i=1}^{n} i = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n = n! \qquad \text{eq. 31.4}$$

Note the fact of 0! = 1. Factorials are used when calculating binomial coefficients for polynomial expansion (Section 31.3) and the number of ways to arrange objects. The number $n!$ states the number of possibilities to arrange n objects (see Sections 1.5.5 and 3.2.3 on entropy and partition functions, respectively).

31.2 QUADRATIC EQUATION

The two solutions of a quadratic equation of the form

$$a \cdot x^2 + b \cdot x + c = 0 \qquad \text{eq. 31.5}$$

are given by

$$x_{1/2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \qquad \text{eq. 31.6}$$

Usually only one of the solutions makes physical sense. A cubic equation of the form

$$a \cdot x^3 + b \cdot x^2 + c \cdot x + d = 0 \qquad \text{eq. 31.7}$$

can also be solved analytically, but needs distinguishing of several cases. These are described in Box 21.10.

31.3 BINOMIAL COEFFICIENTS

Polynomial expansion is often useful in calculus and combinatorics. The binomial theorem states that a non-negative integer power n of a polynomial such as $a + b$ can be expanded into a sum:

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j} \cdot a^j \cdot b^{n-j} = \sum_{j=0}^{n} \binom{n}{j} \cdot a^{n-j} \cdot b^j \qquad \text{eq. 31.8}$$

The expansion is symmetric, it does not matter how we assign a and b. The binomial expansion is a Taylor expansion (Section 31.14.1) with a finite number of n+1 terms. The coefficient for the j^{th} term is calculated using the "n choose j" formula:

$$\binom{n}{j} = \frac{n!}{j!(n-j)!} \qquad \text{eq. 31.9}$$

Expansion of the binomial power

$$(x+1)^n = \sum_{j=0}^{n} \binom{n}{j} \cdot x^j = 1 + n \cdot x + \frac{n(n-1)}{2!} \cdot x^2 + \frac{n(n-1)(n-2)}{3!} \cdot x^3 \ldots \qquad \text{eq. 31.10}$$

followed by rearrangement of the coefficients in n rows of j columns with j ranging from zero to n gives Pascal's triangle (Figure 31.1). The coefficients of the row j+1 are the sums of the coefficients to the left and right of row j.

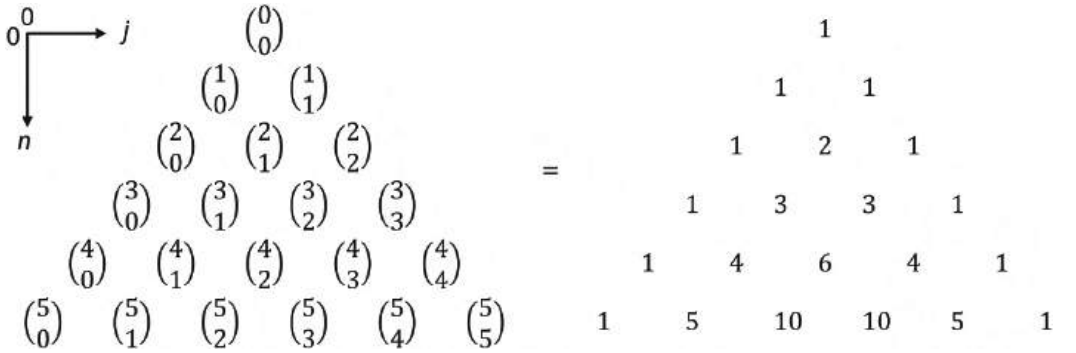


FIGURE 31.1: Pascal's triangle. Triangular array of binomial coefficients.

The relative intensity distributions of isotopologs in mass spectrometry (Section 29.1.5.8), the number of entropic microstates (Section 1.5.5), and the number of species in enzymes with several active sites (Section 13.4) all follow the binomial theorem.

A variant of eq. 31.10 that is used in statistical thermodynamics to describe multiple binding equilibria (Section 3.4) has the control variable j as a factor in the sum. This sum expands to

$$\sum_{j=0}^{n} j \cdot \binom{n}{j} \cdot x^j = n \cdot x \cdot (x+1)^{n-1} \qquad \text{eq. 31.11}$$

31.4 TRIGONOMETRY

In a rectangular triangle, the following rules apply:

$$\sin\beta = \frac{b}{c} \qquad \cos\beta = \frac{a}{c} \qquad \tan\beta = \frac{b}{a} = \frac{\sin\beta}{\cos\beta} \qquad \text{eq. 31.12}$$

For non-rectangular triangles the sine and cosine rule apply:

sine rule:

$$\frac{a}{\sin\alpha} = \frac{b}{\sin\beta} = \frac{c}{\sin\gamma} \qquad \text{eq. 31.13}$$

cosine rule:

$$a^2 = b^2 + c^2 - 2bc\cos\gamma$$
$$b^2 = a^2 + c^2 - 2ac\cos\beta \qquad \text{eq. 31.14}$$
$$c^2 = a^2 + b^2 - 2ab\cos\alpha$$

Thus, if $\gamma = 90°$, we obtain the Pythagorean theorem $a^2+b^2=c^2$ as a special case of the cosine rule. The angles in a triangle add to 180°, and for the two unique angles these conversions can be useful:

$$\sin\alpha \pm \sin\beta = 2\sin\frac{1}{2}(\alpha \pm \beta)\cos\frac{1}{2}(\alpha \mp \beta) \qquad \text{eq. 31.15}$$

$$\cos\alpha + \cos\beta = 2\cos\frac{1}{2}(\alpha + \beta)\cos\frac{1}{2}(\alpha - \beta) \qquad \text{eq. 31.16}$$

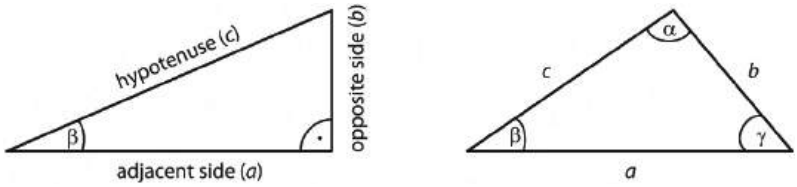


FIGURE 31.2: Rectangular and irregular triangle. Angles are often named arbitrarily, but if the sides of the triangle have letters, the angle opposite to the side is named with the Greek symbol of the side. Thus, the angle opposite the *b*-side is β.

31.5 LOGARITHMS AND EXPONENTIALS

Exponentials and their inverse functions, the logarithms, are ubiquitous in nature. Logarithms are particularly useful to linearize and thus compress large value ranges. The pH and pK_A notation (see Section 2.3) is based on the decadic logarithm, while many thermodynamic functions contain the natural logarithm.

Some useful operations on exponential terms include:

$$e^{\left(a^b\right)} \neq e^{(a \cdot b)} = \left(e^a\right)^b \qquad \text{eq. 31.17}$$

$$e^a \cdot e^b = e^{(a+b)} \qquad \text{eq. 31.18}$$

$$\frac{e^a}{e^b} = e^a \cdot \left(e^b\right)^{-1} = e^a \cdot e^{-b} = e^{(a-b)} \qquad \text{eq. 31.19}$$

The operation given in eq. 31.19 follows directly from a combination of eq. 31.17 and eq. 31.18. The base of the exponential function must be the same for these operations to work. The two most frequently encountered bases for exponentials in nature are Euler's number *e* and ten. In the rare case one encounters exponentials of different bases but with the same exponent, the following relation holds:

$$a^x \cdot b^x = \left(a \cdot b\right)^x \qquad \text{eq. 31.20}$$

Logarithms and exponential functions are inverse functions to one another. The graphs of inverse functions are related to one another by mirror symmetry along the first bisectrix $y = x$. Thus, the natural logarithm $\ln(x)$ is the inverse function of e^x (Figure 31.3). Analogously, the decadic logarithm $\log(x)$, sometimes written $\lg(x)$, is the inverse function of 10^x.

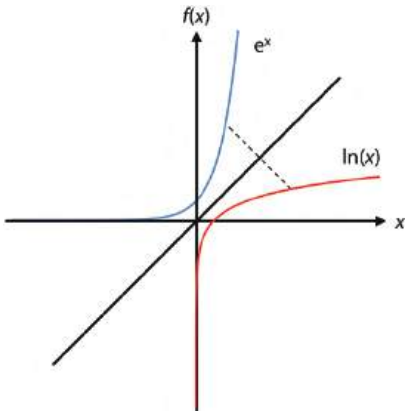


FIGURE 31.3: Inverse functions. A plot of the exponential function e^x and the natural logarithm $\ln(x)$ shows that these functions are related by a mirror operation along the bisecting line $y = x$.

Since exponentials can have any base, so can logarithms. The base is usually just not written, because from the sign *ln* or *log* the base is clear (*e* and 10). For other bases, the logarithm is written as a subscript to *log*, for instance

$$\log_e a = \ln a \qquad \text{eq. 31.21}$$

or

$$\log_{10} a = \log a \qquad \text{eq. 31.22}$$

The inverse operation of logarithms and exponentials can be used to convert the base of any logarithm into another base:

$$\log_b a = \frac{\log_b a}{\log_b b} = \frac{\ln a}{\ln b} = \frac{\log a}{\log b} \qquad \text{eq. 31.23}$$

The term $\log_b b$ equals unity, so if we choose a new base *b*, we can switch between logarithms, for instance

$$\log_{10} a = \frac{\log_{10} a}{\log_{10} 10} = \log a = \frac{\log_e a}{\log_e 10} = \frac{\ln a}{\ln 10} \qquad \text{eq. 31.24}$$

A simple factor of ln10 = 2.303 relates the natural and decadic logarithm. This factor is frequently needed in thermodynamic calculations.

Operations for logarithms are similar to those for exponentials.

1st law: $$\ln a^c = \ln\left(a^c\right) = \ln\left(a\right)^c = c \cdot \ln a \neq \left(\ln a\right)^c \qquad \text{eq. 31.25}$$

It is important that if there are parentheses, they enclose the argument of the logarithm. Potentiation has priority, so if the parentheses enclose the logarithm itself, the exponent cannot be taken as a factor in front of the logarithm.

2nd law: $$\ln a + \ln b = \ln\left(a \cdot b\right) \qquad \text{eq. 31.26}$$

3rd law: $$\ln a - \ln b = \ln\left(\frac{a}{b}\right) \qquad \text{eq. 31.27}$$

The third law is a consequence of the first two:

$$-\ln b = \ln\left(b^{-1}\right) = \ln\frac{1}{b} \qquad \text{eq. 31.28}$$

As a consequence of eq. 31.27 and eq. 31.28, we have

$$\ln\left(\frac{a}{b}\right) = -\ln\left(\frac{b}{a}\right) \qquad \text{eq. 31.29}$$

and

$$\sum_i \ln\left(x_i\right) = \ln\prod_i x_i \qquad \text{eq. 31.30}$$

31.6 DIFFERENTIATION AND INTEGRATION

The first derivative of a function $f(x)$ of a single variable x is written as $\mathrm{d}f(x)/\mathrm{d}x$ or $f'(x)$, the second derivative is written as $\mathrm{d}^2f(x)/\mathrm{d}x^2$, *etc.* The first derivative defines the slope of $f(x)$ at the point $(x|f(x))$. Setting the first derivative to zero allows calculation of the *maxima* ($\mathrm{d}^2f(x)/\mathrm{d}x^2 < 0$) and *minima* ($\mathrm{d}^2f(x)/\mathrm{d}x^2 > 0$) for the function. If the second derivative $\mathrm{d}^2f(x)/\mathrm{d}x^2 = 0$ and $\mathrm{d}^3f(x)/\mathrm{d}x^3 \neq 0$ the function has an *inflection point* at $(x|f(x))$. A few helpful rules are reiterated in the following.

Differentiation of the product of two functions:

product:
$$\frac{\mathrm{d}\left(u(x)\cdot v(x)\right)}{\mathrm{d}x} = u'(x)\cdot v(x) + u(x)\cdot v'(x) \qquad \text{eq. 31.31}$$

Differentiation of the ratio of two functions:

ratio:
$$\frac{\mathrm{d}\left(\dfrac{u(x)}{v(x)}\right)}{\mathrm{d}x} = \frac{u'(x)\cdot v(x) - u(x)\cdot v'(x)}{\left(v(x)\right)^2} \qquad \text{eq. 31.32}$$

A composite of two functions, denoted $u(x)\circ v(x)$, is differentiated using the chain rule:

composition:
$$\frac{\mathrm{d}\left(u(x)\circ v(x)\right)}{\mathrm{d}x} = u'\left(v(x)\right)\cdot v'(x) \qquad \text{eq. 31.33}$$

The rule is used as many times as necessary with functions that have more than two composites. As an example, the function

$$f(x) = e^{\cos x^2} \qquad \text{eq. 31.34}$$

is a triple composite of e^u, $u = \cos v$, and $v = x^2$ and therefore differentiates to

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}u}e^u \cdot \frac{\mathrm{d}}{\mathrm{d}v}\cos v \cdot \frac{\mathrm{d}}{\mathrm{d}x}\left(x^2\right) = e^{\cos x^2}\cdot\left(-\sin x^2\right)\cdot(2x) \qquad \text{eq. 31.35}$$

Sometimes, a function $y = f(x)$ that is not easy to differentiate can be differentiated *via* its inverse function $x = g(y) = f^{-1}(y)$. Since the first derivative of $\mathrm{d}f(x)/\mathrm{d}x$ represents the slope, the first derivative of the inverse function $\mathrm{d}(f^{-1}(y)/\mathrm{d}x = \mathrm{d}(g(y)/\mathrm{d}x$ is the inverse of that slope (see Figure 31.3).

$$g'(y) = \frac{1}{f'(x)} \qquad \text{eq. 31.36}$$

or

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{1}{\mathrm{d}\left(f^{-1}(y)\right)/\mathrm{d}y} \qquad \text{eq. 31.37}$$

For example, the first derivative of $y = f(x) = \tan^{-1}(x) = \arctan(x)$ is not obvious, but the inverse function $x = \tan(y) = \sin(y)/\cos(y)$ is differentiated using eq. 31.32 to

$$\frac{\mathrm{d}\tan(y)}{\mathrm{d}y} = \frac{1}{\cos^2(y)} = \tan^2(y) + 1 \qquad \text{eq. 31.38}$$

The right-hand side of eq. 31.38 uses the relation $\sin^2 y + \cos^2 y = 1$, divided by $\cos^2 y$. Replacing $\tan(y)$ with $\arctan(x)$ and taking the inverse yields the end result:

$$\frac{\mathrm{d}}{\mathrm{d}x}\operatorname{arc\,tan}(x)=\frac{1}{x^2+1} \qquad \text{eq. 31.39}$$

31.6.1 Implicit derivatives

To differentiate a function with respect to a variable, we normally isolate that variable and then take the derivative, also called the explicit derivative. The examples above are all explicit derivatives. However, taking the explicit derivative $\mathrm{d}y(x)/\mathrm{d}x = y'$ of the relation

$$x=y(x)+y^n(x) \qquad \text{eq. 31.40}$$

would require rearrangement of eq. 31.40 to a form $y(x)$ =…, which is impossible. This relation is similar to eq. 18.20 in the discussion on micelle formation (Section 18.4.3). Implicit derivatives can help differentiating any function $f(x,y(x))$ or relation $f(x,y(x)) = g(x,y(x))$ where it is inconvenient or impossible to isolate $y(x)$ to take the explicit derivative. The derivative $\mathrm{d}/\mathrm{d}x$ is taken on both sides of the equation, taking the chain rule into account. We remember that y is a function of x and write $y = y(x)$ for short:

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x,y)=\frac{\mathrm{d}f}{\mathrm{d}x}=\frac{\mathrm{d}f}{\mathrm{d}y}\cdot\frac{\mathrm{d}y}{\mathrm{d}x}=\frac{\mathrm{d}f}{\mathrm{d}y}\cdot y' \qquad \text{eq. 31.41}$$

So we take the derivative of f with respect to y and then multiply with the derivative of y. For the example in eq. 31.40, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}x=\frac{\mathrm{d}}{\mathrm{d}x}\left(y+y^n\right) \qquad \text{eq. 31.42}$$

The left-hand side is the derivative of x, which is 1. The right-hand side can be split into the individual terms of the sum:

$$1=\frac{\mathrm{d}}{\mathrm{d}x}y+\frac{\mathrm{d}}{\mathrm{d}x}y^n \qquad \text{eq. 31.43}$$

The first term on the right is just the derivative of y, or y'. Applying the chain rule to the second term results in

$$1=y'+n\cdot y^{n-1}\cdot\frac{\mathrm{d}}{\mathrm{d}x}y=y'+n\cdot y^{n-1}\cdot y' \qquad \text{eq. 31.44}$$

We then isolate y' for the result:

$$y'=\frac{1}{1+n\cdot y^{n-1}} \qquad \text{eq. 31.45}$$

31.6.2 Partial and total derivatives

Functions with two or more variables, *e.g.* $y = f(x_1,x_2)$, can have partial or total differentials. The partial differential $\partial/\partial x_i$ for variable x_i is obtained when differentiating the function for x_i while treating the others as constants:

$$\left(\frac{\partial f(x_1,x_2)}{\partial x_1}\right)_{x_2} \qquad \text{eq. 31.46}$$

The total differential is the sum of all partial differentials.

$$\frac{dy}{dx_i} = \left(\frac{\partial f(x_1,x_2)}{\partial x_1}\right)_{x_2} + \left(\frac{\partial f(x_1,x_2)}{\partial x_2}\right)_{x_1} \qquad \text{eq. 31.47}$$

Note that the letter "d" signifies an infinitesimal change, while the letter ∂ indicates the partial differential. A further source of confusion is the Greek letter δ, which signifies changes in a state function (Section 1.2).

31.6.3 Integration

Integration of a function yields the area under the curve. General rules for integration include:

factoring out constants: $$\int c \cdot f(x)dx = c \cdot \int f(x)dx \qquad \text{eq. 31.48}$$

rule of sums: $$\int \left(u(x)+v(x)\right)dx = \int u(x)dx + \int v(x)dx \qquad \text{eq. 31.49}$$

product rule: $$\int \left(u'(x)\cdot v(x)\right)dx = u(x)\cdot v(x) - \int u(x)\cdot v'(x)\cdot dx \qquad \text{eq. 31.50}$$

substitution rule: $$\int u\left(v(x)\right)dx = \int u(w)\frac{dw}{w'} \qquad \text{eq. 31.51}$$

where $v(x) = w$ is a clever choice of substitution that results in a product that can be integrated. The integration variable changes from dx to dw according to $w' = dw/dx$. For example, the following function can be integrated by substituting $w = (ax+b)$ and dx = dw/a:

$$\int f(x)dx = \int \frac{1}{ax+b}dx = \int \frac{1}{w}\frac{dw}{a} = \frac{1}{a}\int \frac{dw}{w} = \frac{1}{a}\cdot \ln w = \frac{\ln(ax+b)}{a} \qquad \text{eq. 31.52}$$

In the last step the substitution is reverted.

Some useful integrals and derivatives are shown in Table 31.1. The integration constants have been omitted for clarity.

TABLE 31.1:
Useful integrals and differentials.

$\frac{df(f)}{dx}$	$f(x)$	$F(x)=\int F(x)dx$
0	$n\in\mathbb{R}$	$n\cdot x$
$n\cdot x^{n-1}$	x^n	$\frac{1}{n+1}\cdot x^{n+1}$
$n\cdot e^{nx}$	e^{nx}	$\frac{1}{n}e^{nx}$
$\ln n\cdot n^x$	$n^x \quad (n>0)$	$\frac{n^x}{\ln n}$
$\frac{n}{x}$	$n\cdot\ln x$	$n\cdot x\cdot\ln x-n\cdot x$
$x^{n-1}\cdot(n\cdot\ln x+1)$	$x^n\cdot\ln x \quad (n\geq 0)$	$\frac{x^{n+1}}{n+1}\cdot\left(\ln x-\frac{1}{n+1}\right)$
$n\cdot\cos(nx)$	$\sin(nx)$	$-\frac{\cos(nx)}{n}$
$\cosh x$	$\sinh x$	$\cosh x$
$2\cdot\sin x\cdot\cos x$	$\sin^2 x$	$\frac{1}{2}\cdot(x-\sin x\cdot\cos x)$
$-2\cdot\sin x\cdot\cos x$	$\cos^2 x$	$\frac{1}{2}\cdot(x+\sin x\cdot\cos x)$
$1+\tan^2 x$	$\tan x$	$-\ln\lvert\cos x\rvert$
$n\cdot\cos(2nx)$	$\sin nx\cdot\cos nx$	$-\frac{\cos^2(nx)}{2n}$
$\frac{u''(x)\cdot u(x)-(u'(x))^2}{(u(x))^2}$	$\frac{u'(x)}{u(x)}$	$\ln\lvert u(x)\rvert$
$(u(x))^n\cdot\left(u''(x)+n\cdot\frac{(u'(x))^2}{u(x)}\right)$	$u'(x)\cdot(u(x))^n$	$\frac{1}{n+1}\cdot(u(x))^{n+1}$

31.6.4 Differential equations

Differential equations contain both a function and one or more of its derivatives. The highest derivative determines the order of the differential equation. A second-order differential equation therefore has the form $y = f(x) = k_1 \cdot d/dx + k_2 \cdot d^2/dx^2$. Kinetic rate laws (Chapters 6 and 7) are derived by establishing the differential equations for the change of concentrations of the reactants. On several occasions (eq. 11.14, eq. 11.32, eq. 21.95) in this book we encounter differential equations of the type

$$a\cdot\frac{dx(t)}{dt}=c-b\cdot x(t) \qquad \text{eq. 31.53}$$

These can be solved either by separation of variables or as linear equations. Separation of variables with the shortcut $x(t) = x$ gives

$$a\cdot\frac{dx}{c-b\cdot x}=dt \qquad \text{eq. 31.54}$$

Substitution of $u = c–b·x$ for the denominator and using

$$\frac{\mathrm{d}u}{\mathrm{d}x} = -b \qquad \text{eq. 31.55}$$

gives

$$a \cdot \frac{\mathrm{d}u}{-b \cdot u} = \mathrm{d}t \qquad \text{eq. 31.56}$$

This integrates to

$$-\frac{a}{b} \cdot \int \frac{\mathrm{d}u}{u} = -\frac{a}{b} \cdot \int \mathrm{d}\ln u = \int_{t_0}^{t} \mathrm{d}t \qquad \text{eq. 31.57}$$

or

$$-\frac{a}{b} \cdot \ln u = t - t_0 \qquad \text{eq. 31.58}$$

With $t_0 = 0$, substituting back u gives

$$-\frac{a}{b} \cdot \left[\ln\left(c - b \cdot x\right)\right]_{x_0}^{x} = -\frac{a}{b} \cdot \left[\ln\left(c - b \cdot x\right) - \ln\left(c - b \cdot x_0\right)\right] = t \qquad \text{eq. 31.59}$$

If $x_0 = 0$, then

$$-\frac{a}{b} \cdot \left(\ln \frac{\left(c - b \cdot x\right)}{c}\right) = t \qquad \text{eq. 31.60}$$

Solving for x results in

$$x = \frac{c}{b} \cdot \left(1 - e^{-\frac{b}{a} \cdot t}\right) \qquad \text{eq. 31.61}$$

The same result is obtained when treating the problem as a linear equation. We re-write eq. 31.53 as

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} + \frac{b}{a} \cdot x(t) = \frac{c}{a} \qquad \text{eq. 31.62}$$

and multiply both sides with $\exp(bt/a)$:

$$\frac{\mathrm{d}x}{\mathrm{d}t} \cdot e^{\frac{bt}{a}} + x \cdot \frac{b}{a} \cdot e^{\frac{bt}{a}} = \frac{c}{a} \cdot e^{\frac{bt}{a}} \qquad \text{eq. 31.63}$$

We realize that

$$\frac{b}{a} \cdot e^{\frac{bt}{a}} = \frac{d}{\mathrm{d}t} \cdot e^{\frac{bt}{a}} \qquad \text{eq. 31.64}$$

and hence eq. 31.63 becomes

$$\frac{\mathrm{d}x}{\mathrm{d}t} \cdot e^{\frac{bt}{a}} + x \cdot \frac{\mathrm{d}}{\mathrm{d}t} \cdot e^{\frac{bt}{a}} = \frac{c}{a} \cdot e^{\frac{bt}{a}} \qquad \text{eq. 31.65}$$

The right-hand side of eq. 31.65 is the result of differentiation by the product rule (eq. 31.31):

$$\frac{\mathrm{d}x}{\mathrm{d}t}\cdot e^{\frac{bt}{a}}+x\cdot\frac{\mathrm{d}}{\mathrm{d}t}\cdot e^{\frac{bt}{a}}=\frac{\mathrm{d}}{\mathrm{d}t}\cdot\left(x\cdot e^{\frac{bt}{a}}\right) \qquad \text{eq. 31.66}$$

The reverse product rule is used for finding the integral:

$$\int\frac{\mathrm{d}}{\mathrm{d}t}\cdot\left(x\cdot e^{\frac{bt}{a}}\right)\cdot\mathrm{d}t=\frac{c}{a}\cdot\int e^{\frac{bt}{a}}\cdot\mathrm{d}t \qquad \text{eq. 31.67}$$

or

$$\left[x\cdot e^{\frac{bt}{a}}\right]_{x_0}^{x}=\frac{c}{b}\cdot\left[e^{\frac{bt}{a}}\right]_{t_0}^{t} \qquad \text{eq. 31.68}$$

For $x_0 = 0$, and $t_0 = 0$, evaluation results in

$$x\cdot e^{\frac{bt}{a}}=\frac{c}{b}\cdot\left(e^{\frac{bt}{a}}-1\right) \qquad \text{eq. 31.69}$$

Solving for x yields eq. 31.61.

31.7 PARTIAL FRACTIONS

Expansion into partial fractions is useful to simplify a rational function before integrating it (see eq. 7.33 in Chapter 7 on integrated rate laws). The method is based on the fact that any rational function can be expressed as the sum of a polynomial function $P(x)$ and another rational function $Q(x)$:

$$P(x)+Q(x)=\sum_{i=1}^{n}a_i\cdot x^i+\frac{Z(x)}{N(x)} \qquad \text{eq. 31.70}$$

where $Z(x)$ is of lower degree than $N(x)$. Arriving at this form may require polynomial long division. $Q(x)$ is of the form

$$Q(x)=\frac{Z(x)}{N(x)}=\sum_{j=1}^{m}\frac{c_j}{(x-x_j)^k} \qquad \text{eq. 31.71}$$

The numbers x_j in the rational function are the poles, where $Q(x)$ reaches infinity. They are also the zeros of multiplicity k of $N(x)$. If the zeros of $N(x)$ are known, expansion into partial fractions reduces to finding the coefficients c_j. An example that is used in the kinetics part of this book is the function:

$$\frac{1}{\left(x^2-2xA_0B_0+A_0{}^2B_0{}^2\right)}=\frac{1}{(A_0-x)(B_0-x)}=\frac{c_1}{(A_0-x)}+\frac{c_2}{(B_0-x)} \qquad \text{eq. 31.72}$$

The poles are A_0 and B_0 and $Z(x) = 1$. $Q(x)$ should be written as a sum of ratios with the same denominators in the form shown to right of eq. 31.72, which can then be integrated without much difficulty. Multiplying eq. 31.72 with the denominator $(A_0-x)(B_0-x)$, and rearrangement yields

$$1=c_1\cdot(B_0-x)+c_2\cdot(A_0-x)=-(c_1+c_2)\cdot x+c_1B_0+c_2A_0 \qquad \text{eq. 31.73}$$

Equating the coefficients in eq. 31.73 requires that

$$1=c_1B_0+c_2A_0 \qquad \text{eq. 31.74}$$

and

$$0 = -(c_1 + c_2) \cdot x \qquad \text{eq. 31.75}$$

This system of linear equations leads to $c_1 = -c_2$ and

$$c_1 = -c_2 = -\frac{1}{A_0 - B_0} \qquad \text{eq. 31.76}$$

Going back to eq. 31.72 then yields the end result:

$$\frac{1}{(A_0 - x)(B_0 - x)} = \frac{1}{A_0 - B_0} \cdot \left(\frac{1}{(B_0 - x)} - \frac{1}{(A_0 - x)} \right) \qquad \text{eq. 31.77}$$

31.8 L'HÔPITAL'S RULE

This rule is useful for finding the limes or limiting value of a function that is the ratio of two other functions and where both of these functions approach zero or diverge to infinity. An example is the function

$$f(x) = \frac{g(x)}{h(x)} = \frac{\sin(x)}{x} \qquad \text{eq. 31.78}$$

The function $f(x)$ is undefined for $x = 0$, but the limiting value of $f(x)$ approaches unity for $x \to 0$. Simply using the limes for $g(x)$ and $h(x)$, as is usually done, is not helpful because both are zero.

$$\lim_{x \to 0} f(x) = \frac{\lim_{x \to 0} g(x)}{\lim_{x \to 0} h(x)} = \frac{0}{0} \qquad \text{eq. 31.79}$$

l'Hôpital's rule states that the limes of the function $f(x)$ is the same as the ratio of the limes of the derivatives of the component functions:

$$\lim_{x \to x_0} f(x) = \frac{\lim_{x \to x_0} \frac{\mathrm{d}}{\mathrm{d}x} g(x)}{\lim_{x \to x_0} \frac{\mathrm{d}}{\mathrm{d}x} h(x)} \qquad \text{eq. 31.80}$$

In the $\sin(x)/x$ example above this leads to

$$\lim_{x \to x_0} \frac{\sin(x)}{x} = \frac{\lim_{x \to x_0} \frac{\mathrm{d}}{\mathrm{d}x} \sin(x)}{\lim_{x \to x_0} \frac{\mathrm{d}}{\mathrm{d}x} (x)} = \frac{\lim_{x \to x_0} \cos(x)}{\lim_{x \to x_0} 1} \qquad \text{eq. 31.81}$$

The cosine function approaches unity for $x \to 0$.

31.9 VECTORS

A vector is a quantity with a defined magnitude and direction, for example velocity. In the three-dimensional Euclidian space that we live in, three Cartesian coordinates describe a point. The coordinates can be viewed as the tip of a vector with a magnitude that is the shortest distance between the point (x,y,z) and the origin of the coordinate system. The vector $\boldsymbol{u}$ then points from the origin $(0,0,0)$ to (x,y,z). Using velocity as an example, the three coordinates are the component movements along the x-, y-, and z-directions that contribute to the overall movement. In molecular dynamics

calculations (Section 20.1.5), a vector sum (the resulting force) is calculated for all individual forces acting simultaneously on an atom due to its interactions with other atoms. Vectors are usually written in boldface to distinguish them from numbers.

$$\boldsymbol{u} = (x, y, z) = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u_1\boldsymbol{i} + u_2\boldsymbol{j} + u_3\boldsymbol{k} \qquad \text{eq. 31.82}$$

The components u_i of the vector can be multiplied by the orthogonal unit vectors $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ along the x-, y-, and z-axes to write the vector $\boldsymbol{u}$ as a *vector sum*. It does not matter if we denote the coordinates of the vector horizontally or vertically, or with subscripts, as long as we use a consistent notation. Vectors can be translated without changing their properties, which is useful to illustrate vector addition and subtraction (and is used in the Patterson map; Figure 24.20). Vectors are added and subtracted from one another by operating on the individual elements:

$$\boldsymbol{u} - \boldsymbol{v} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} - \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = \begin{pmatrix} u_1 - v_1 \\ u_2 - v_2 \\ u_3 - v_3 \end{pmatrix} \qquad \text{eq. 31.83}$$

In geometric terms, vector addition results in another vector (Figure 31.4).

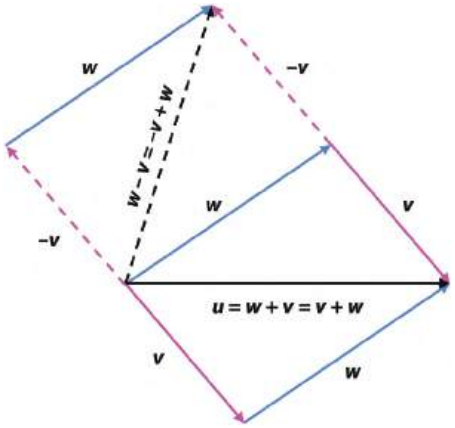


FIGURE 31.4: Vector addition and subtraction. Vectors in opposite directions are drawn as dashed lines. The operations are commutative, *i.e.* the order of the vectors for addition and subtraction is irrelevant.

The magnitude of a vector is given by the Pythagoras formula

$$|\boldsymbol{u}| = \|\boldsymbol{u}\| = \sqrt{u_1^2 + u_2^2 + u_3^2} \qquad \text{eq. 31.84}$$

The magnitude of vectors is sometimes stated between double vertical bars to distinguish it from an absolute value of a number.

31.9.1 Dot product

Angles between two vectors, *e.g.* $\boldsymbol{u}$ and $\boldsymbol{v}$, are calculated using the dot product ($\boldsymbol{u}\cdot\boldsymbol{v}$), which is also known as the *scalar product* because the result of the operation is a scalar. The dot product is defined as the product of the vector magnitudes times the cosine of the angle θ that the vectors enclose:

$$(\boldsymbol{u}\cdot\boldsymbol{v}) = u_1v_1 + u_2v_2 + u_3v_3 = |\boldsymbol{u}||\boldsymbol{v}|\cos\theta \qquad \text{eq. 31.85}$$

From the coordinates of three bonded atoms, two vectors are calculated and the 1,3-angle or bond angle between the vectors may be calculated according to eq. 31.85.

31.9.2 Cross product

The result of the cross product or *vector product* is another vector. This vector ($\boldsymbol{u}\times\boldsymbol{v}$) is perpendicular to the plane that contains the original vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and its magnitude is the area of the parallelogram that is spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$:

$$(\boldsymbol{u}\times\boldsymbol{v}) = (u_2v_3 - v_2u_3)\boldsymbol{i} + (u_1v_3 - v_1u_3)\boldsymbol{j} + (u_1v_2 - v_1u_2)\boldsymbol{k} \qquad \text{eq. 31.86}$$

Cross products are used in the calculation of torsion angles (Section 17.2.1.2). A torsion or dihedral angle between four bonded atoms 1–2–3–4 may be defined as the angle between the vectors normal to the planes spanned by atoms 1–2–3 and 2–3–4. The three unit vectors $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ in eq. 31.86 form a right-handed system, which means that the cross product is not commutative, or $(\boldsymbol{u}\times\boldsymbol{v})\neq(\boldsymbol{v}\times\boldsymbol{u})$. Swapping the order of the vectors in a cross product results in a vector of opposite direction. A *scalar triple product* $\boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w})$ is the scalar product of $\boldsymbol{u}$ with the vector product of $\boldsymbol{v}$ and $\boldsymbol{w}$. It is the volume of the parallelepiped spanned by the three vectors $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$, *e.g.* the unit cell volume of a crystal (eq. 24.3).

31.10 MATRICES

Matrices are used to structure and simplify mathematical objects such as systems of equations. They are rectangular arrays of numbers or expressions, distributed over m rows and n columns. The expression $m\times n$ is the *dimensionality* of the matrix, spoken "m-by-n". The numbers or expressions in a matrix M are called the *entries* a_{ij} with i running from row 1…m and j running from column 1…n. Eq. 31.87 shows the general form of a matrix. The dimensionality $m\times n$ when viewed as a product reports the total number of entries in the matrix.

$$M = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{nm} \end{pmatrix} \qquad \text{eq. 31.87}$$

For example, a 3×4 matrix has dimensionality 12. To illustrate how matrices can help simplify a mathematical object, consider the shorthand notation for a linear system of equations

$$\begin{aligned} a\cdot x_1 + b\cdot x_2 &= c \\ d\cdot x_1 + e\cdot x_2 &= f \end{aligned} \qquad \text{eq. 31.88}$$

which becomes

$$\begin{pmatrix} a & b \\ d & e \end{pmatrix}\cdot\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ d & e \end{pmatrix}\cdot\vec{x} = \begin{pmatrix} c \\ f \end{pmatrix} \qquad \text{eq. 31.89}$$

Here, the matrix $\begin{pmatrix} a & b \\ d & e \end{pmatrix}$ is a square (2×2) matrix because the number of columns is equal to the number of rows, or $n = m = 2$. It also shows that the a_{ij} notation may be dropped for small matrices to allow for simpler reading. $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and $\begin{pmatrix} c \\ f \end{pmatrix}$ in eq. 31.89 are called *column vectors*, they are the usual vectors we encounter in this book and may also be written with an arrow on top or in boldface (Section 31.9). Accordingly, individual columns of the matrix such as $\begin{pmatrix} a \\ d \end{pmatrix}$ and $\begin{pmatrix} b \\ e \end{pmatrix}$ in eq. 31.89 may

also be regarded as column vectors. *Row vectors* are just the individual rows of the matrix, or (*a b*) and (*d e*) in eq. 31.89. In other words, a vector may be viewed as a simple matrix with either $n = 1$ or $m = 1$. In a square matrix ($n = m$), the diagonal from the top left to the bottom right has the entries $a_{11}, a_{22}, \ldots a_{nn}$ and the sum of this set of entries is known as the *trace* of the matrix.

31.10.1 Basic operations on matrices

Operations on matrices, often done in the background without the user noticing, are done whenever an object is moved or a system of equations is solved. Object movements are discussed in molecular modeling (Section 20), in X-ray crystallography (Section 24; *e.g.* in crystal symmetry and molecular replacement), and in cryo-EM (Section 26.5; *e.g.* image alignment and three-dimensional reconstruction). Systems of equations are solved when deriving integrated rate laws (Section 9).

Matrices may be multiplied by a number (scalar) or with a vector. They may be added, subtracted, and transposed. Scalar multiplication of a matrix M by a number n means multiplying each element a_{ij} of the matrix with n:

$$n \cdot M = n \cdot \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} n \cdot a_{11} & n \cdot a_{12} \\ n \cdot a_{21} & n \cdot a_{22} \end{pmatrix} \qquad \text{eq. 31.90}$$

Addition and subtraction of matrices M and N also operate entrywise:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix} \qquad \text{eq. 31.91}$$

Subtraction of matrices is simply multiplication with –1 and addition.

$$M - N = M + (-1) \cdot N \qquad \text{eq. 31.92}$$

It will be clear from eq. 31.91 that addition (and subtraction) work only with matrices of the same dimensionality $m \times n$.

Matrices may be multiplied by a vector, and this is actually the operation converting eq. 31.89 into a form like that shown in eq. 31.88:

$$\begin{pmatrix} a & b \\ d & e \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a \cdot x_1 + b \cdot x_2 \\ d \cdot x_1 + e \cdot x_2 \end{pmatrix} \qquad \text{eq. 31.93}$$

In general, an $m \times n$ matrix M may be multiplied by a column vector $\vec{x}$ that has the same number of entries as there are columns in M, which is n. The resulting product vector $\vec{p}$ will also have n entries.

$$M \cdot \vec{x} = \vec{p} \qquad \text{eq. 31.94}$$

The entries in $\vec{p}$ are the dot product (eq. 31.85) of $\vec{x}$ with the rows in M, which explains why $\vec{x}$ must have the same number of entries as the matrix has columns:

$$M \cdot \vec{x} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11} \cdot x_1 + a_{12} \cdot x_2 + \ldots + a_{1n} \cdot x_n \\ a_{21} \cdot x_1 + a_{22} \cdot x_2 + \ldots + a_{2n} \cdot x_n \\ \vdots \\ a_{m1} \cdot x_1 + a_{m2} \cdot x_2 + \ldots + a_{mn} \cdot x_n \end{pmatrix} \qquad \text{eq. 31.95}$$

The extension of multiplication of a matrix with a vector, which is a single-column matrix, is multiplication of two matrices:

$$M \cdot N = P \qquad \text{eq. 31.96}$$

The same condition applies: the matrix N needs to have the same number of rows as M has columns. If M is $m \times n$ *and* N *is* $n \times p$, then the product matrix P is an $m \times p$ matrix:

$$P = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \cdot \begin{pmatrix} b_{11} & \cdots & b_{1p} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{np} \end{pmatrix} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & \ddots & \vdots \\ c_{m1} & \cdots & c_{mp} \end{pmatrix} \quad \text{eq. 31.97}$$

where each entry c_{ij} in matrix P is the dot product between the i^{th} row in matrix M and the j^{th} column in matrix N, or

$$c_{ij} = a_{i1} \cdot b_{1j} + a_{i2} \cdot b_{2j} + \ldots + a_{in} \cdot b_{nj} \quad \text{eq. 31.98}$$

Note that matrix (and vector) multiplications are not commutative, *i.e.*:

$$M \cdot N \neq N \cdot M \quad \text{eq. 31.99}$$

The transposition of a matrix means swapping the elements from rows and columns, *i.e.* by turning rows into columns and *vice versa*. This operation changes the dimensionality from $m \times n$ to $n \times m$. For example, a 2×3 matrix M will give a 3×2 matrix M^T:

$$M = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}$$

$$M^T = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}^T = \begin{pmatrix} a & c & e \\ b & d & f \end{pmatrix} \quad \text{eq. 31.100}$$

Matrix transposition also often occurs "in the background", *e.g.* in weight estimation and back-propagation of weights through neural networks and in deep learning (Section 20.2). Further examples include image and signal processing where the computer implementation of the discrete Fourier transform requires matrix transposition.

A special matrix that becomes important for solving systems of differential equations is the *identity matrix I*, which is a square matrix with values of unity along the main diagonal from top left to bottom right and zero elsewhere, for example in this 3×3 identity matrix of trace = 3:

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{eq. 31.101}$$

The identity matrix is useful to allow operations between matrices and scalars, as we will see in the calculation of Eigenvalues.

31.10.2 Solving systems of linear equations by Gaussian elimination

A system of linear equations encoded by a matrix may be solved using *Gaussian elimination*, also known as the row-reduction method. This algorithm uses three operations to convert a matrix into an *upper triangular matrix*. The upper triangular matrix is a square matrix that has zeros in its lower left-hand corner (below its main diagonal). The operations allowed are swapping of two rows, multiplying a row by a non-zero number, and adding a multiple of one row to another row. Using the system of two linear equations in eq. 31.89, the upper triangular matrix would be

$$\begin{pmatrix} a & b \\ 0 & ae - bd \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c \\ af - cd \end{pmatrix} \quad \text{eq. 31.102}$$

The form is achieved by multiplying the first row in eq. 31.89 with $-d$, multiplying the second row by a, and adding the two rows to eliminate the entry at the lower left (the variable d). The solutions of the linear system of equations are now apparent from the upper triangular matrix:

$$(ae-bd)\cdot x_2 = af-cd \qquad \text{eq. 31.103}$$

$$x_2 = \frac{af-cd}{ae-bd} \qquad \text{eq. 31.104}$$

Inserting this result for x_2 into the equation given by the first row, we obtain a solution for x_1:

$$a\cdot x_1 + b\cdot x_2 = c \qquad \text{eq. 31.105}$$

$$a\cdot x_1 + b\cdot \frac{af-cd}{ae-bd} = c \qquad \text{eq. 31.106}$$

$$a\cdot x_1 = c - b\cdot \frac{af-cd}{ae-bd} \qquad \text{eq. 31.107}$$

$$x_1 = \frac{c-b\cdot\frac{af-cd}{ae-bd}}{a} = \frac{ce-bf}{ae-bd} \qquad \text{eq. 31.108}$$

With larger matrices the Gaussian elimination method may become unwieldy. An alternative method to solving such systems is the use of Eigenvalues and Eigenvectors.

31.10.3 Eigenvalues and Eigenvectors to solve systems of linearly independent equations

Eigenvalues λ and Eigenvectors $\vec{\eta}$ are fundamental properties of square matrices M that help in solving systems of equations. Consider the product of the $n\times n$ matrix M and the $n\times 1$ matrix (a vector) $\vec{\eta}$, which gives another vector $\vec{y}$:

$$M\cdot\vec{\eta} = \vec{y} \qquad \text{eq. 31.109}$$

We can now ask if there is a special vector $\vec{\eta}$ that gives as a result

$$\vec{y} = \lambda\cdot\vec{\eta} \qquad \text{eq. 31.110}$$

In other words, we ask that the result of matrix multiplication with a vector should yield that same vector multiplied with a constant factor λ:

$$M\cdot\vec{\eta} = \lambda\cdot\vec{\eta} \qquad \text{eq. 31.111}$$

Rearranging,

$$M\cdot\vec{\eta} - \lambda\cdot\vec{\eta} = \vec{0} \qquad \text{eq. 31.112}$$

introducing the identity matrix,

$$M\cdot\vec{\eta} - \lambda\cdot I\cdot\vec{\eta} = \vec{0} \qquad \text{eq. 31.113}$$

and factoring out $\vec{\eta}$ gives

$$\left(M-\lambda\cdot I\right)\cdot\vec{\eta}=\vec{0} \qquad \text{eq. 31.114}$$

For all matrices M such constant(s) λ and vector(s) $\vec{\eta}$ do exist, and are termed the *Eigenvalue*(s) and *Eigenvectors* of matrix M. There may be a single Eigenvalue or several Eigenvalues, and they can be real or complex numbers. Each Eigenvalue has infinitely many Eigenvectors associated with it, which are, however, all linearly dependent. The requirement for the solution of a system of differential equations reduces to finding the Eigenvalues and Eigenvectors for that system. A trivial and physically not meaningful solution to eq. 31.114 is $\vec{\eta}=\vec{0}$, which we do not consider further. For $\vec{\eta}\neq\vec{0}$, this leaves the condition:

$$\left(M-\lambda\cdot I\right)=\vec{0} \qquad \text{eq. 31.115}$$

The right-hand side of eq. 31.115 will be the zero vector only if the matrix is singular, *i.e.* if its determinant is zero:

$$det\left(M-\lambda\cdot I\right)=0 \qquad \text{eq. 31.116}$$

The determinant of a 2×2 matrix is given by

$$det\begin{pmatrix}a_1 & b_1\\ a_2 & b_2\end{pmatrix}=\begin{vmatrix}a_1 & b_1\\ a_2 & b_2\end{vmatrix}=a_1b_2-a_2b_1 \qquad \text{eq. 31.117}$$

and that of a 3×3 matrix is given by

$$\begin{vmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ a_3 & b_3 & c_3\end{vmatrix}=a_1b_2c_3+a_2b_3c_1+a_3b_1c_2-a_3b_2c_1-a_2b_1c_3-a_1b_3c_2 \qquad \text{eq. 31.118}$$

Determinants for larger matrices are tedious to calculate, but online calculators exist. A 4×4 matrix involves calculation of the determinants of four 3×3 matrices:

$$\begin{vmatrix}a_1 & b_1 & c_1 & d_1\\ a_2 & b_2 & c_2 & d_2\\ a_3 & b_3 & c_3 & d_3\\ a_4 & b_4 & c_4 & d_4\end{vmatrix}= \qquad \text{eq. 31.119}$$

$$a_1\cdot\begin{vmatrix}b_2 & c_2 & d_2\\ b_3 & c_3 & d_3\\ b_4 & c_4 & d_4\end{vmatrix}-a_2\cdot\begin{vmatrix}b_1 & c_1 & d_1\\ b_3 & c_3 & d_3\\ b_4 & c_4 & d_4\end{vmatrix}+a_3\cdot\begin{vmatrix}b_1 & c_1 & d_1\\ b_2 & c_2 & d_2\\ b_4 & c_4 & d_4\end{vmatrix}-a_4\cdot\begin{vmatrix}b_1 & c_1 & d_1\\ b_2 & c_2 & d_2\\ b_3 & c_3 & d_3\end{vmatrix}$$

For the case of a 2×2 matrix, the equation one needs to solve in order to obtain the Eigenvalues λ is

$$det\left(M-\lambda\cdot I\right)=det\begin{pmatrix}a_1-\lambda & b_1\\ a_2 & b_2-\lambda\end{pmatrix}=\left(a_1-\lambda\right)\left(b_2-\lambda\right)-a_2b_1=0 \qquad \text{eq. 31.120}$$

Likewise, for a 3×3 matrix, the equation one needs to solve for λ is

$$\left|M-\lambda\cdot I\right|=\begin{vmatrix} a_1-\lambda & b_1 & c_1 \\ a_2 & b_2-\lambda & c_2 \\ a_3 & b_3 & c_3-\lambda \end{vmatrix}=0 \qquad \text{eq. 31.121}$$

The degree of the polynomial to be solved equals the dimensionality of the matrix: the 2×2 matrix potentially has two Eigenvalues as the solutions of a quadratic equation. A 3×3 matrix with up to three Eigenvalues requires solving a cubic equation. In general, an $n\times n$ matrix results in an n^{th} degree polynomial, termed the *characteristic polynomial*. The sum of all Eigenvalues equals the trace of the matrix M. Once we derived the Eigenvalues λ from the characteristic polynomial, we can go back to eq. 31.114 and determine the associated Eigenvectors $\vec{\eta}$.

31.11 COMPLEX NUMBERS

A complex number has two parts, termed *real* (a) and *imaginary* ($i\cdot b$) (eq. 31.122). These terms are just naming conventions for two properties of a quantity that need to be kept separated. In this respect, the word "imaginary" is misleading.

$$z=a+i\cdot b$$

$$i=\sqrt{-1} \qquad \text{eq. 31.122}$$

$$i^2=-1$$

The separator is i, defined as the square root of –1. Complex numbers allow equations to be solved that have no real solutions, *e.g.* $z^2-2z+2=0$, which has the complex solutions $z_{1/2}=1\pm i$. A graphical representation of complex numbers is the *Argand diagram* or *complex plane* (Figure 31.5).

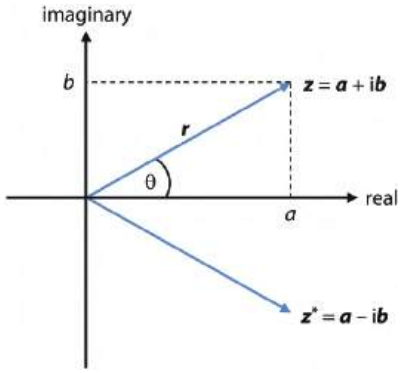


FIGURE 31.5: Argand diagram. Two equivalent representations of z are possible, one using Cartesian notation a and b, the other polar notation r and θ, where r is the magnitude and θ is the phase angle. If θ is a function of time t, the arrow will rotate at a frequency ωt in the ab-plane. z and z^* are complex conjugates.

The components of the point z can also be expressed by polar coordinates using the magnitude r and the angle θ:

$$a=r\cdot\cos\theta$$

$$b=r\cdot\sin\theta \qquad \text{eq. 31.123}$$

This leads to

$$z=a+i\cdot b=r\cdot\left(\cos\theta+i\cdot\sin\theta\right) \qquad \text{eq. 31.124}$$

With the help of the Euler formula eq. 31.125, complex numbers in the polar coordinate notation can be written in a more compact form that is also mathematically simpler to process:

$$e^{i\theta} = \cos\theta + i \cdot \sin\theta \qquad \text{eq. 31.125}$$

Substituting $-\theta$ for θ in eq. 31.125 gives an equivalent Euler formula:

$$e^{-i\theta} = \cos\theta - i \cdot \sin\theta \qquad \text{eq. 31.126}$$

For the special case of $\theta = 2\pi$, the complex number becomes real:

$$e^{\pm i \cdot 2\pi} = \cos 2\pi \pm i \cdot \sin 2\pi = 1 \qquad \text{eq. 31.127}$$

Similarly:

$$e^{\pm i \cdot \pi} = -1$$
$$e^{\pm i \cdot \frac{\pi}{2}} = \pm i \qquad \text{eq. 31.128}$$

Adding and subtracting equations eq. 31.125 and eq. 31.126 re-defines the sine and cosine functions as complex exponentials:

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}$$
$$\sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} \qquad \text{eq. 31.129}$$

Two complex numbers that just differ by the sign of the complex part are *complex conjugates*, termed z^* (Figure 31.5). In other words, the complex conjugate can be found by replacing all instances of i with $-i$:

$$z^* = \left(a + i \cdot b\right)^* = \left(a - i \cdot b\right)$$
$$z^* = r \cdot \left(\cos\theta - i \cdot \sin\theta\right) \qquad \text{eq. 31.130}$$
$$z^* = r \cdot e^{-i\theta}$$

The square root of the product of a complex number with its conjugate is the *magnitude*, *absolute value*, or *modulus* of the complex number:

$$|z| = \sqrt{z \cdot z^*}$$
$$|z| = \sqrt{\left(a + i \cdot b\right) \cdot \left(a - i \cdot b\right)} = \sqrt{a^2 + b^2} = r \qquad \text{eq. 31.131}$$

The same result of $|z| = r$ is obtained for the other two expressions of the complex conjugate in eq. 31.130. The magnitude of a complex number is equivalent to r in the Argand diagram (Figure 31.5).

Arithmetic operations are carried out similar to vectors. For two complex numbers $x = a+ib$ and $y = c+id$

$$x+y=(a+c)+i\cdot(b+d)$$

$$x-y=(a-c)+i\cdot(b-d)$$

$$x\cdot y=(ac-bd)+i\cdot(ad+bc)$$

eq. 31.132

$$\frac{x}{y}=\frac{a+ib}{c+id}=\frac{a+ib}{c+id}\cdot\frac{c-id}{c-id}=\frac{(ac+bd)}{c^2+d^2}+i\cdot\frac{(bc-ad)}{c^2+d^2}$$

When dividing complex numbers, the complex conjugate is used to make the denominator real. Multiplication and division of complex numbers is easier in their exponential form because the standard rules described for exponentials in Section 31.5 apply:

$$r_1\cdot e^{i\theta_1}\cdot r_2\cdot e^{i\theta_2}=r_1\cdot r_2\cdot e^{i(\theta_1+\theta_2)}$$

$$\frac{r_1\cdot e^{i\theta_1}}{r_2\cdot e^{i\theta_2}}=\frac{r_1}{r_2}\cdot e^{i(\theta_1-\theta_2)}$$

eq. 31.133

The Euler formula describes a periodic function with a repeat of $i\cdot 2\pi$. This can be understood by taking $\theta_2 = 2\pi$, as in eq. 31.127 above, and adding this angle to θ_1:

$$e^{i\theta_1}\cdot e^{i\cdot 2\pi}=e^{i(\theta_1+2\pi)}=e^{i\theta_1}$$

eq. 31.134

The ease of angle addition, which equals multiplication of complex exponentials, is one of the reasons why harmonic wave functions are often expressed as complex exponentials (see Sections 24.4.2 and 24.4.4) rather than sine and cosine waves. Both the real ($r\cdot\cos\theta$) and imaginary ($r\cdot\sin\theta$) parts can be chosen to describe a harmonic wave function. In practice, calculations are done with the complex form, and later the real part is extracted from the solution to represent the actual wave. In addition, the observable quantity, *e.g.* in scattering and diffraction, is the intensity. The intensity is proportional to the square of the amplitude (modulus or absolute value) of the light wave, which is always a real number.

31.12 BASIC ELEMENTS OF STATISTICS

Statistics provides the tools to analyze numerical data. One of them is the *arithmetic mean* $\langle x\rangle$ of a set of N data values x_i, which is defined as the sum of all values divided by the total number of values:

$$\langle x\rangle=\bar{x}=\mu=\frac{1}{N}\cdot\sum_{i=1}^{N}x_i$$

eq. 31.135

The arithmetic mean, also written as Greek letter μ or with an overstrike as $\bar{x}$, is an approximation for the true value of the variable x. Whenever data are averaged (*e.g.* absorption, fluorescence, X-ray reflection intensities, NMR FIDs, projections in EM, *etc.*), the aim is to increase the signal-to noise ratio and obtain a more accurate estimate for the true value of x.

A sometimes more robust estimate for the center of the dataset is the *median* value, which is the number half way along a sorted list of data points. The median is less influenced by extreme outliers than the mean. For instance, the set (1,2,3,4,100) has a mean of 22, which is larger than most data points, and a median of 3, which is closer to most data points. In this case, an outlier rejection would be appropriate. If the number N of data points is uneven, the median is taken as the mean of the two points closest to $N/2$. Figure 20.3 gives an example of the use of the median.

A measure of how closely the data points cluster around the mean is the *mean squared error* or *variance* σ^2, which is the average of the squared deviations from the mean:

$$\sigma^2 = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \bar{x}^2 \qquad \text{eq. 31.136}$$

The square is important because deviations may be positive or negative. When summed up without squaring, these deviations may cancel each other, possibly giving the wrong impression of a small overall deviation when there really are large individual deviations with opposite signs. Squaring before summation, however, changes the unit of the variables x_i, which is solved by taking the square root of the variance. The positive square root of the variance is the *standard deviation* σ:

$$\sigma = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x} \right)^2} \qquad \text{eq. 31.137}$$

Note that N in eq. 31.137 changes to $(N–1)$ if the set of data values analyzed is only a fraction of the whole dataset and thus serves to *estimate* the standard deviation for the whole dataset. If $N–1$ is used, σ is named the *experimental standard deviation*. Likewise, the variances σ^2 are known as *population variance* for N data points and *sample variance* for a subset of the population, where $N–1$ is taken in eq. 31.136.

For large numbers of N, it does not matter whether N or $N–1$ are used. If $N \geq 50$, the deviation between the two ways to calculate σ is $\leq 1\%$. Variances, errors, and standard deviations are used in many places in this book, *e.g.* for the calculation of correlation coefficients (eq. 24.13), in energy functions (eq. 24.26), in estimating the focal spread of electron beams, and as part of Gaussian functions (eq. 31.141), *e.g.* in fluorescence intensity distributions (Section 21.5.9.1).

Standard deviation and mean are combined into the *standard score*, or *Z-score*, which gives the number of standard deviations of a data point above or below the mean of the data:

$$Z = \frac{x - \bar{x}}{\sigma} \qquad \text{eq. 31.138}$$

Positive and negative Z-scores are used in the display of difference electron density during crystallographic model building (Section 24.4.5). A frequently asked question is how reliable is the mean value $\langle x \rangle$ of a dataset, what is its standard deviation? This quantity is known as the *standard error* that spans a *confidence interval* around the variable x. The standard error may be calculated from the mean and standard deviation of the means of several measured datasets. Alternatively, if only a single dataset is available, the standard error is estimated from the standard deviation σ of that dataset by:

$$\sigma_{\langle x \rangle} = \frac{\sigma}{\sqrt{N}} \qquad \text{eq. 31.139}$$

For example, in order to reduce the error on the standard deviation by a factor of two, four times as many data points are required. If we measure infinitely many data points N for the variable x, we would get its *true value*, which in reality is unknown. However, the confidence interval defined by the standard error σ means that if we measure variable x many times, in 68% of all cases it will lie within that interval. A doubling or tripling of the interval to 2σ or 3σ increases the probability of finding the true value of x within the larger interval to 95% and 99.7%, respectively (Figure 31.6).

If the values in a dataset are sorted into bins from small to large, and the relative frequency of the values (number of values in a bin divided by total number of values in the dataset) is plotted as a function of the average of the values in the bin, a bell-shaped curve is often obtained. Such data values show a *normal distribution* or *Gaussian distribution* (Figure 31.6), which for an infinitely large number of data points in infinitely narrow bins follows the general formula:

$$f(x) = a \cdot e^{-\frac{x^2}{b}} \qquad \text{eq. 31.140}$$

Specifically, the probability density distribution is defined as

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} \qquad \text{eq. 31.141}$$

where the maximum of the function is the mean $\bar{x}$, also known as the expectation value. The standard deviation σ is the distance between the inflection points and the expectation value.

When integrated from $-\infty$ to $+\infty$, this function has an area of 1. The shape of the distribution is strongly influenced by the variance. Large variances (values scattered widely around the mean) give a broad distribution while small variances give a narrow distribution.

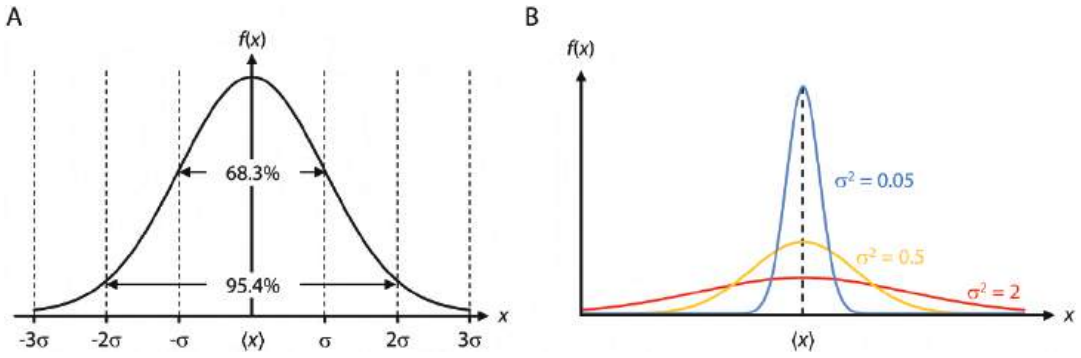


FIGURE 31.6: Normal distribution. Left: 68.3% of all data points are within $\langle x \rangle \pm \sigma$, 95.4% are within $\langle x \rangle \pm 2\sigma$, and 99.7% are within $\langle x \rangle \pm 3\sigma$ (not shown). Right: The functions have the same mean and the same area (they are normalized) but different variances. Large values of σ^2 give a broader distribution.

For an array of data, the *correlation coefficient CC* can establish if a correlation exists and, if so, how good it is. A two-dimensional array of (x,y) number pairs has a *CC* of

$$CC = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2}} \qquad \text{eq. 31.142}$$

CC values range from $-1 \leq CC \leq 1$. While for $CC = 0$ there is no correlation, positive and negative values for *CC* mean correlation and anti-correlation, respectively, between the values in the pairs (x,y). *CC* values are used to gauge the practical resolution limit of X-ray diffraction (Section 24.4.3) and cryo-EM data (Section 26.5.3).

31.13 ERROR PROPAGATION

There are basically three kinds of errors, gross, systematic, and random, which can be time-dependent or not. Gross errors such as defective equipment or inappropriate signal for the process to be measured should be avoided because they cannot be corrected for later. Systematic errors can arise due to wrong calibration of instruments, sample inhomogeneity/impurity, disturbing electromagnetic fields, or when exceeding the limits of a physical law (linearity, for example). Here, we treat time-independent, random or statistical errors from sources such as changing environment parameters during the measurement (temperature, pH, voltage), quantum effects (noise, fluctuations), or simply errors when manually recording measurement values. When several data points or data series, with their associated errors, are combined into a new quantity, linear or Gaussian error

propagation is needed to calculate the error of that quantity. Which of the two methods should be used is not always clear-cut, but in general depends on how the data combining into the new quantity were measured. For single-point data, where the errors are estimated directly from the measurement, linear error propagation is appropriate. If the error sources are purely random, as is often observed, Gaussian error propagation is preferable. If the new quantity is combined from data having both types of errors, either linear error propagation or a combination with Gaussian error propagation is used.

Linear error propagation for a function $x = f(a, b, c,\ldots)$ that depends on the independent variables $a, b, c\ldots$ with the associated errors $\sigma_a, \sigma_b, \sigma_c\ldots$, estimates the overall linear error σ_x by

$$\sigma_x = \sigma_a \cdot \left|\frac{\partial x}{\partial a}\right| + \sigma_b \cdot \left|\frac{\partial x}{\partial b}\right| + \sigma_c \cdot \left|\frac{\partial x}{\partial c}\right| \cdots \qquad \text{eq. 31.143}$$

The equation above is a Taylor expansion (see next Section 31.14.1) of the function x that is canceled after the first term. The operators $\partial/\partial a$, $\partial/\partial b$, and $\partial/\partial c$ denote the partial derivatives of the function x with respect to a, b, and c. The partial derivatives are weighted by the estimated maximum errors σ and taken as absolute values such that positive errors in one variable and negative errors in another cannot compensate each other (errors are always positively additive).

In Gaussian error propagation, for the same function $x = f(a, b, c,\ldots)$, the overall error σ_x is estimated by the variance formula as

$$\sigma_x = \sqrt{\sigma_a^2 \cdot \left(\frac{\partial x}{\partial a}\right)^2 + \sigma_b^2 \cdot \left(\frac{\partial x}{\partial b}\right)^2 + \sigma_c^2 \cdot \left(\frac{\partial x}{\partial c}\right)^2 \cdots} \qquad \text{eq. 31.144}$$

where the squared errors $\sigma_{a,b,c}$ are the variances of the variables a, b, and c. The variance formula assumes that the errors $\sigma_{a,b,c}$ are uncorrelated, purely statistical, and follow a Gaussian distribution.

If $x = a \pm m$ where m is a constant, the error for variable a in both types of propagation does not change and is independent of m:

$$\sigma_{a\pm m} = \sigma_a \cdot \left|\frac{\partial x}{\partial a}\right| = \sigma_a \cdot 1 = \sigma_a \qquad \text{eq. 31.145}$$

$$\sigma_{a\pm m} = \sqrt{\sigma_a^2 \cdot \left(\frac{\partial x}{\partial a}\right)^2} = \sqrt{\sigma_a^2 \cdot 1^2} = \sigma_a \qquad \text{eq. 31.146}$$

If, however, $x = n{\cdot}a \pm m{\cdot}b$, the error is weighted by the factors n and m, but in different ways:

$$\sigma_x = \sigma_a \cdot \left|\frac{\partial x}{\partial a}\right| + \sigma_b \cdot \left|\frac{\partial x}{\partial b}\right| = \sigma_a \cdot n + \sigma_b \cdot m \qquad \text{eq. 31.147}$$

$$\sigma_x = \sqrt{\sigma_a^2 \cdot \left(\frac{\partial x}{\partial a}\right)^2 + \sigma_b^2 \cdot \left(\frac{\partial x}{\partial b}\right)^2} = \sqrt{n^2\sigma_a^2 + m^2\sigma_b^2} \qquad \text{eq. 31.148}$$

For example, given measured values for the free energy ΔG and the enthalpy ΔH, from $\Delta G = \Delta H – T\,\Delta S$ the error of the entropy change ΔS is calculated as

$$\sigma_{\Delta S} = \sigma_{\Delta H} \cdot \frac{1}{T} + \sigma_{\Delta G} \cdot \left|-\frac{1}{T}\right| = \frac{1}{T}\left(\sigma_{\Delta H} + \sigma_{\Delta G}\right) \qquad \text{eq. 31.149}$$

$$\sigma_{\Delta S} = \sqrt{\sigma_{\Delta H}^2 \cdot \left(\frac{1}{T}\right)^2 + \sigma_{\Delta G}^2 \cdot \left(-\frac{1}{T}\right)^2} = \frac{1}{T}\sqrt{\sigma_{\Delta H}^2 + \sigma_{\Delta G}^2} \qquad \text{eq. 31.150}$$

As a further example, if we have a series of intensity measurements I with associated errors σ_I and would like to convert the I into $\ln(I)$, the error is propagated to

$$\sigma_{\ln I} = \sqrt{\sigma_I^2 \cdot \left(\frac{\partial \ln(I)}{\partial I}\right)^2} = \sqrt{\sigma_I^2 \cdot \left(\frac{1}{I}\right)^2} = \frac{\sigma_I}{I} \qquad \text{eq. 31.151}$$

For such data, the errors are statistical and thus only Gaussian error propagation is appropriate. A case where error propagation is essential to recognize is the linearized forms of equations describing enzymatic steady-state reactions (Box 13.1).

31.14 SERIES EXPANSION

31.14.1 Taylor series

Taylor series expansion is used to approximate a continuous function $f(x)$ about a point $(x_0|f(x_0))$. The function must be infinitely differentiable and single-valued.

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(x_0)}{n!} \cdot (x - x_0)^n$$

$$= f(x_0) + f'(x_0) \cdot (x - x_0) + \frac{f''(x_0)}{2!} \cdot (x - x_0)^2 + \frac{f'''(x_0)}{3!} \cdot (x - x_0)^3 + \ldots \qquad \text{eq. 31.152}$$

If we stop the series expansion after $n = 1$, we obtain a linear function, which is the tangent to the function $f(x)$ at the point $(x_0|f(x_0))$. While the tangent is not a good approximation for the function, the parabolic function obtained after the third term is often already good enough and is used in some energy minimization algorithms. The more terms we include in the expansion, the better will the power series describe the function $f(x)$ also far away from x_0. Taylor series expansion is used in the Newton-Raphson energy minimization algorithm (Section 20.1.2) and to approximate Newtonian equations of motion (Section 20.1.5.2).

For $x_0 = 0$ the Taylor series is known as a *Maclaurin series*. For instance, the Maclaurin series for the exponential function $f(x) = e^x$ is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots \qquad \text{eq. 31.153}$$

and the sine function expands to

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n \cdot x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots \qquad \text{eq. 31.154}$$

In order for the series to converge, $|x| < 1$.

A number of trigonometric relations can be verified using the Maclaurin series:

$$\sin^2 x + \cos^2 x = 1$$

$$\cos(x_1 + x_2) = \cos x_1 \cdot \cos x_2 - \sin x_1 \cdot \sin x_2 \qquad \text{eq. 31.155}$$

$$\cos x_1 + \cos x_2 = 2 \cdot \cos\frac{x_1 + x_2}{2} \cdot \cos\frac{x_1 - x_2}{2}$$

The Euler formula relates the exponential function to the sine and cosine functions, as can be shown by Maclaurin expansion. This form of the exponential function is used to evaluate Fourier transformations for electron density calculations (Section 24.4.2):

$$e^{ix} = \cos x + i \cdot \sin x$$

$$e^{-ix} = \cos x - i \cdot \sin x \qquad \text{eq. 31.156}$$

31.14.2 Fourier series

In 1807, the French physicist Joseph Fourier introduced the concept that every function $f(x)$ of periodicity $2L$ (frequency $2\pi\nu_0 = \omega_0 = \pi/L$) can be expanded into an infinite series of sine and cosine waves of frequency $\omega = n\pi/L$, now called its *Fourier series*:

$$f(x) = \frac{1}{2}a_o + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L} + \sum_{n=1}^{\infty} b_n \sin\frac{n\pi x}{L} \qquad \text{eq. 31.157}$$

The coefficients of the series are

$$a_0 = \frac{1}{L}\int_0^{2L} f(x)\mathrm{d}x$$

$$a_n = \frac{1}{L}\int_0^{2L} f(x)\cos\frac{n\pi x}{L}\mathrm{d}x \qquad \text{eq. 31.158}$$

$$b_n = \frac{1}{L}\int_0^{2L} f(x)\sin\frac{n\pi x}{L}\mathrm{d}x$$

Using the Euler formula, the Fourier series can also be written in complex form

$$f(x) = \sum_{n=0}^{\infty} c_n e^{in\omega_o x}\mathrm{d}x \qquad \text{eq. 31.159}$$

with the Fourier coefficients c_n:

$$c_n = \frac{1}{L}\int_0^{2L} f(x)e^{in\omega_o x}\mathrm{d}x \qquad \text{eq. 31.160}$$

To illustrate the application of the Fourier series, a *square wave* of length $2L$ in the time domain that we would like to approximate is plotted in Figure 31.7. Its Fourier series expansion has only sine terms:

$$\frac{4}{\pi}\cdot\sum_{n=1,3,5\ldots}^{\infty}\frac{1}{n}\cdot\sin\frac{n\pi t}{L} \qquad \text{eq. 31.161}$$

Since the number $n\pi/L$ corresponds to the frequency of the sine term n, later terms in the series correspond to higher frequencies and thus higher resolutions.

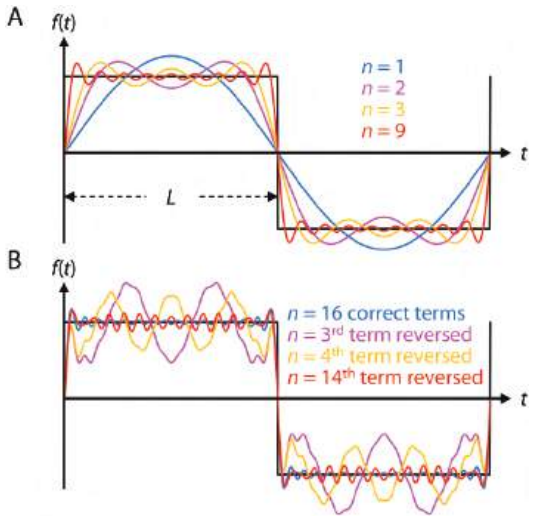


FIGURE 31.7: Fourier series. The square wave is shown in black. A: The first approximation is a sine wave (blue) corresponding to only one term ($n = 1$). More sine wave terms are summed in the other functions, and in the progress ever better approximate the square function. B: Effect of wrong phase on the Fourier series. The blue curve is the Fourier series with 16 correct terms. Phase reversal of a single term, from sin(nx) to sin($-nx$) = –sin(nx), strongly affects the Fourier series.

From Figure 31.7, we see that more terms better describe the function. In practice, as many terms as are needed or can be computed in a certain amount of time are used for such calculations. It can also be seen that near the corners of the square function a "ringing" or spike remains in the Fourier series that will not go away with larger numbers of terms. This *Gibbs phenomenon* is an overshoot of the Fourier series, typical near points of discontinuity, and can lead to artifacts in electron density maps. More worrisome, however, are errors in phases (Section 24.4.5). Reversal of the phase in just a single term results in strong changes of the whole Fourier series. As expected, phase errors in early terms ($n = 3$ or $n = 4$ in Figure 31.7) have a stronger effect than phase errors for higher n. Still, phase reversal of a later term ($n = 14$ out of 16 total terms in Figure 31.7) has a noticeable effect on the shape of the series.

31.15 FOURIER TRANSFORMATION

The Fourier transform (FT) is an extension of the Fourier series concept to include non-periodic functions (Figure 31.8). As the period of the function $f(t)$ increases, an increasing number of frequency terms is needed to describe it. For periods approaching infinity, where the function is essentially non-periodic, the frequency spectrum is continuous, *i.e.* has an infinite number of frequency terms. The units of time-dependent amplitude A transform into units multiplied by time (symbolized by $A{\cdot}s$).

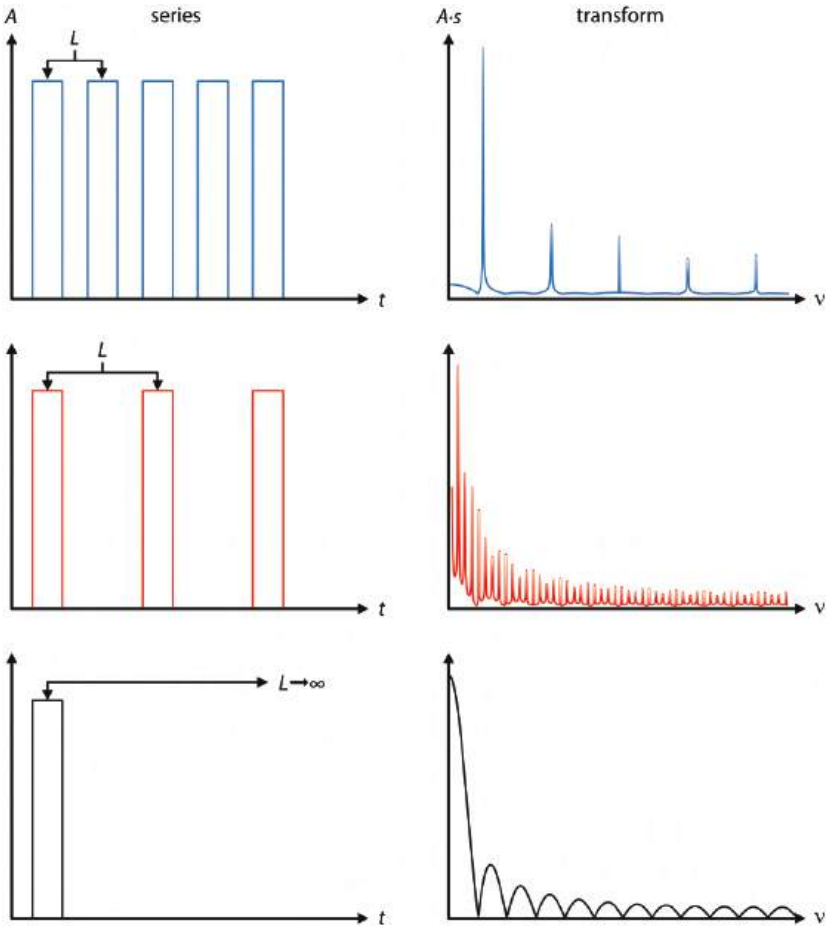


FIGURE 31.8: Fourier series and Fourier transforms. Periodic functions (left) and their Fourier transforms (right). The smaller the period, the fewer peaks there are in the Fourier transform and the farther apart they are spaced. For infinite periods, the Fourier transform spectrum is continuous.

The relationship between a time-domain function $f(t)$ and its frequency domain function $F(\nu)$ is

$$f(t) = \int_{-\infty}^{\infty} F(\nu) e^{2\pi i \nu t} \mathrm{d}\nu \qquad \text{eq. 31.162}$$

and the inverse Fourier transform is

$$F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \nu t} \mathrm{d}t \qquad \text{eq. 31.163}$$

Since $f(t)$ is usually not continuous but a list of N measurement values taken at small regular time intervals δt, a *discrete Fourier transform* is needed that uses summation rather than integration.

$$F(k\delta\nu) = \delta t \sum_{n=0}^{N-1} f(n\delta t) e^{-\frac{2\pi i n k}{N}} \qquad \text{eq. 31.164}$$

The function $f(n\delta t)$ expresses the experimental values, and the result is a set of N frequency values $F(k\delta\nu)$ or *Fourier coefficients*. These coefficients are separated by $\delta\nu = 1/N\delta t$, showing the reciprocal relationship established by the Fourier transformation. To calculate each $F(k\delta\nu)$, eq. 31.164 has to be evaluated for each value of k, so the computational time needed for the Fourier transform scales

with N^2, a severe limitation for large datasets (*e.g.* diffraction data, Sections 24.1, 24.4.2, or NMR FIDs, Section 22.1.4). The introduction of the fast Fourier transform (FFT) algorithm, which scales with $N \cdot \ln N$, has therefore had a huge impact on scientific computation. The inverse operation from frequencies to time data using the discrete Fourier transform is analogous:

$$f(n\delta t) = \frac{1}{N}\sum_{k=0}^{N-1} F(k\delta\nu)e^{\frac{2\pi i n k}{N}} \qquad \text{eq. 31.165}$$

Figure 31.9 shows two classical Fourier transforms that are often encountered in biophysics. The Fourier transform of a square function (just one period of a square wave) is of the form $\sin(\nu)/\nu$. Square functions describe light pulses (Figure 22.15) or diffraction slits. The measured responses oscillate with a frequency that is inversely correlated to the pulse length or the slit width. Another example is the exponentially damped oscillation, such as the FID in NMR spectroscopy (Figure 22.16). The Fourier transform of this function is a *Lorentzian function* that has both real and imaginary components. The real part is the *absorption spectrum*, and the imaginary part is the *dispersion spectrum*. The line width of the Lorentzian function is proportional to the damping constant of the exponential decay, *i.e.* the faster the decay of the signal, the broader the absorption and dispersion spectra.

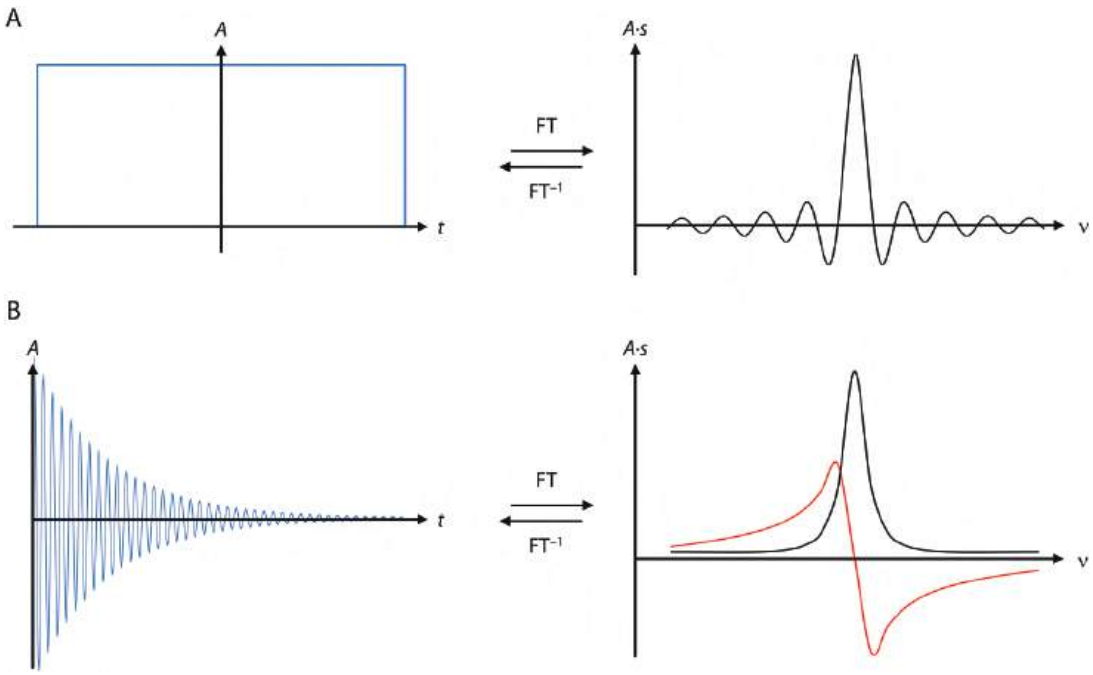


FIGURE 31.9: Example Fourier transforms. A: The Fourier transform of a square function (pulse of amplitude *A*) in the time domain transforms into frequency. For light diffraction at a slit, the slit width *x* will transform into a frequency spectrum with the distance between the diffraction maxima proportional to 1/*x*. B: The Fourier transform of an exponentially damped oscillation yields a Lorentzian function with real (black) and imaginary (red) spectral components.

31.16 CONVOLUTION

A convolution (sometimes called German: Faltung) is the integral that reports the amount of overlap of two functions as one function, g(*t*), is shifted across the other function, *f*(*t*):

$$f(t) \otimes g(t) = \int_{-\infty}^{\infty} f(\tau)\cdot g(t-\tau)\tau = \int_{-\infty}^{\infty} g(\tau)\cdot f(t-\tau)\mathrm{d}\tau \qquad \text{eq. 31.166}$$

The convolution therefore blends one function with another function. Its operator symbol is the asterisk (*) or the tensor product ⊗. The connection of convolution to Fourier transformation is the *convolution theorem*, which states that the Fourier transform of a convolution of functions $f(t)$ and $g(t)$ is the product of the individual FTs $F(\nu)$ and $G(\nu)$, and *vice versa*, that the Fourier transform of the product of two functions is the convolution of their FTs:

$$FT\left(f(t)\otimes g(t)\right)=F(\nu)\cdot G(\nu)$$
$$FT\left(f(t)\cdot g(t)\right)=F(\nu)\otimes G(\nu) \qquad \text{eq. 31.167}$$

Thus, convolution in the time domain, for example in the measurement of fluorescence lifetimes (Section 21.5.9.1), is equivalent to pointwise multiplication in the frequency domain. An important property of the convolution is that the convolution of two Gaussians is another Gaussian.

Convolution is also a central aspect in crystallography. Mathematically speaking, a crystal is the convolution of an object (the content of the unit cell) with a crystal lattice (a symmetry function). The diffraction pattern of a crystal is therefore the product of the diffraction patterns of the lattice (defining the *hkl* triplets) and the unit cell (defining the intensities). The *deconvolution* of the observed diffraction pattern separates the symmetry aspect from the unit cell content. Other examples where convolutions are observed include the Patterson function (24.4.3), the contrast transfer function in TEM (Section 26.4.3), and ESI mass spectra (Section 29.1.4).

Index

0

0th law of thermodynamics, 6

1

180° pulse, 581
1-anilinonaphtalene-8-sulfonate (ANS), 529
1st law of thermodynamics, 7

2

2-aminopurine, 522, 529, 549
2D classification, 759
2nd law of thermodynamics, 22, 28, 30

3

3_{10}-helix, 332
3D classes, 761
3rd law of thermodynamics, 33

9

90° pulse, 581

A

ab initio method, 305, 456, 609, 634
Abbe limit, 692, 724
Aberration (of lens), 737
absolute entropy, 33
absolute zero, 9
absorption, 475
  of X-rays, 669
  probability, 493
  to monitor structural changes, 490
acceleration, 449, 504, 827
accessible surface area, 348
acetaldehyde (NMR spectrum), 576
acetyl-CoA, 288
acid-base catalysis, 278, 311
action potential, 138
activated complex, 266
activation barrier, 266, 270, 373
activation energy, 191, 266, 268
  in crystallization, 642
  in enzyme catalysis, 276
activation enthalpy, entropy, and free energy, 268
active transport, 118
active-site titration, 525
activity coefficient, 57, 69
Adair equation, 102
adenylate kinase, 352
adiabatic, 3
  compression, 15
  expansion, 31
  expansion and compression, 15, 20
  process, 18
affinity label, 249
A-form (of DNA), 409, 411
A-form (of RNA), 428
aggregation number, 357
Airy disc, 691
aldehyde coupling, 316, 843
alignment
  (multiple) sequence- and structure-based, 325
  of dipoles, 299
  of nuclear spins, 572
  of protein domains, 344
allosteric inhibition, 237, 353
allowed transitions, 476
alpha (α)-helix, 330
AlphaFold2, 456
  in molecular replacement, 672
alternating laser excitation (ALEX), 722
Alzheimer's disease, 388
amide proton exchange, 387, 594, 596, *See also* H/D exchange
amine coupling, 842
amino acids, 308
  frequency in proteins, 323
amphiphile
  lipids and detergents, 356
  α-helix, 330
amplitude contrast, 749
amyloid plaques, 388
analytical ultracentrifugation, 820, *See also* ultracentrifugation
Anfinsen, C., 372
Ångstrøm, 651
anisotropy
  of chemical shift (in NMR), 593, 597
  of fluorescence, 537
  time dependence, 546
  to follow binding events, 540, 541, 791
anomalous diffraction, 667
anomer, 289, 401
anthracyclines, 427
antiparallel β-sheet, 334

antiport, 118
apparent diffusion coefficient, 705
apparent equilibrium constant, 76, 215, 379
apparent rate constant, 164, 166, 386
Archimedes' principle, 823
Argand diagram, 660, 666, 893
argon (polarizability of), 303
Arrhenius equation, 266
ascorbic acid, 316, 388, *See* vitamin C
asparagine glycosylation, 314
association
  constant, 72, 166
  in sedimentation equilibrium, 833
astigmatism, 737
asymmetric stretching vibration, 507
asymmetric unit, 649
  and non-crystallographic symmetry, 651
  versus biological unit, 651
  volume of, 651
atomic force microscopy (AFM), 766
atomic form factor, *See* atomic scattering factor
atomic scattering factor, 625, 660, 668, 741
ATP
  biosynthesis, 116
  fluorescent, 522
  hydrolysis, 118
ATP synthase, 116, 135
attention mechanism, 454, 459
Auger electrons, 647
autocorrelation function, 620, 631, 701, 707
autonomous folding unit, 341
autoprotolysis, 76
Avogadro constant, 36
Avogadro's law, 8
axial coma, 737

B

back focal plane, 747
band centrifugation, 834
bandpass filter, 694
base catalysis, 278
base stacking, 405
  and duplex formation, 407
  energy, 407
base triplet, 412
base-pair geometries, 399, 408
bathochromic effect, 479, 483
bending magnet, 655
beta (β)-barrel, 337, 342, 366
beta (β)-galactosidase, 487
beta (β)-strands and β-sheets, 334
beta (β)-turn, 337
B-form (of DNA), 409
bilayer, 289, *See also* lipid bilayer (membrane)
bimolecular reaction, 147, 155, 165, 268, 797
binomial coefficient, 35, 226, 814, 876
biochemical standard conditions, 44
biochemical standard potential, 130
bioinformatics, 443
biological chromophores, 481, 501
biological fluorophores, 514
biological membranes, 60, 118, 136, 354, 361
biotin-streptavidin affinity, 353
biotin-streptavidin coupling, 435, 522, 708, 717, 777, 844
blinking
  of fluorophores, 709, 722
  of quantum dots, 522
Bohr magneton, 599
boiling point elevation, 58
Boltzmann
  distribution, 87, 89, 474, 578
  equation, 35
  factors, 91
bonds, 290
  angles, 292
  breakage during isomerization, 289
  breakage during isomerization, 288
  energy and length, 292
  stretching and bending, 507
  vibration, 504
  π-bond, 289, 294
  σ-bond, 289
boundary condition, 158, 450, 592
Boyle-Mariotte law, 8
Bragg's law, 658
Bravais lattice, 649
Bremsstrahlung, 653
Briggs-Haldane formalism, 216
brown fat, 118
Brownian diffusion (Brownian motion), 111
Brownian ratchet, 120
buffer capacity, 76
buffer properties, 858
bumble bee, 24
buoyancy
  in ultracentrifugation, 822
  of fish, 9
buried surface area, 348
B-value, 677

C

caged compound, 795
calmodulin, 319, 556, 629
calorimeter, 853
carbon dioxide, 507
carboxylation, 317
Carnot process, 20, 29
carotenes, 483
carrier gas, 804, 810, 816
catalysis, 273
  acid-base, 278, 311
  by antibodies, 277
  by enzymes, 276
  covalent, electrophilic, and nucleophilic, 281
  intramolecular, 282
catenanes (in DNA), 417
cation-π interaction, 299
central slice theorem, 758
chain rule, 880
chain tracing (in NMR), 589
chaperone, 345, 385, 794
charge density wave, 838
charge-charge interactions, 296
charge-coupled device (CCD), 662, 739
charge-dipole interactions, 296, 299
Charles' law, 8
ChatGPT, 454
chemical exchange, 594, 596
chemical potential, 46
  and principle of Le Chatelier, 71
  as partial molar property, 51
  in analytical ultracentifugation, 831
  in crystallization, 642
  of mixtures, 53
  of protons, 116
  of solutions, 55
  of solvent, 58, 61
  pressure and temperature dependence, 51
chemical shift, 572
  anisotropy, 593, 597

definition, 573
of amino acids, 589
perturbation, 574
sensitivity to local environment, 574
chemiosmotic hypothesis, 116
chevron plot, 386
chicken foot structure, 416
chiral center, 289
chlorophyll, 135, 352, 483, 495
Chou-Fasman method, 340
chromatic aberration, 737, 756
chromogenic substrate, 487
chromophore, 477, 481, 501, 624
chymotrypsin, 83, 249, 281
circular birefringence, 497
circular dichroism (CD), 497
of proteins and nucleic acids, 501
circularly polarized light, 496
classical electron radius, 625
Clausius inequality, 30
closed system, 3, 37
coalescence, 594
coarse-grained modeling, 448, 634
codon usage, 404
coefficient
activity, 57, 69
diffusion, 110, 824
extinction, 478
friction, 548, 620, 824
Hill, 231
sedimentation, 820, 827
transmission, 268
coherence
of electrons, 734, 737
of phases in NMR, 581, 585
of X-rays, 681
coiled-coil, 331
cold denaturation, 43, 377
collagen helix, 316, 332, 388
colligative properties, 57
collision theory, 269
collisional quenching, 535
collision-induced decay (CID), 810
colocalization single-molecule spectroscopy (CoSMoS), 716
complex number, 661, 893
Compton scattering, 615
computational biology, 443
concentration dependence
in nucleic acid annealing, 100
of chemical potential μ, 53
of observed rate constants, 202
of sedimentation coefficient, 828
concentration determination
of nucleic acids, 482, 487
of proteins, 485
concentration gradient, 23, 63, 110, 118, 135, 137, 643, 820, 834
concerted model of cooperativity, 229
configuration, 287
enantiomers and diastereomers, 289
confocal microscopy, 695, 723
confocal volume, 620, 695, 701
conformation, 290, *See also* conformers
alternate, 675
anti (in DNA), 411
anti and *gauche*, 294
Newman and sawhorse projections, 293
staggered and eclipsed, 293
conformational energy, 294, 306
conformational selection, 352
conformers, 290
in helical junctions, 720
in SAXS, 635
stabilization of, 560
time scale of interconversion, 382, 558
conjugate acid and base, 75
conjugate gradient algorithm, 446, 676
consecutive reactions, 168
constant
Avogadro, 36
Boltzmann, 35
dielectric, 296
dissociation, 73
equilibrium, 69
Faraday, 116, 126
force, 294, 505, 808
gas, 4
hyperfine coupling, 601
inhibition (K_I) and relation to IC_{50}, 240
ionization (of buffers), 858
Michaelis-Menten, 215
Planck, 268
rate, 146
spring, 291, 292, 808
Wiseman, 856
constitutional isomers, 287
contact mode (in AFM), 767
contour length, 771
contrast reversal, 751
contrast transfer function (CTF), 751
effective, 754
contrast variation, 627, 635, 636, 637
convolution, 544, 903
in diffraction, 659
in electron microscopy, 751
in fluorescence lifetime measurement, 544
in NMR, 584
in Patterson function, 664
cooperativity
of enzymes, 228
of hydrogen bonds, 301
of micelle formation, 357
of nucleic acid annealing, 100
of protein folding, 370, 375
copper (in electrochemistry), 127
corralled diffusion, 114
correlated spectroscopy (COSY), 587
cosine rule, 877
cost function, 454
Coulomb potential, 741
Coulomb's law, 296
coupled enzymatic assay, 488
coupling constant, 575, 601
covalent catalysis, 281
covalent interactions, 290
critical angle, 697, 838
critical micelle concentration (*cmc*), 356
cross product, 888
cross-correlation function, 707
cross-peaks, 587
cross-relaxation, 577
cross-section, 110, 357, 625, 653
cruciform structure (in DNA), 414
cryo-electron tomography (cryo-ET), 756
cryoscope, 60
crystal system, 648
crystallization of macromolecules, 641
crystallographic symmetry, 651
crystallography, 640
electron, 682
kinetic, 679
neutron, 682
serial, 680

crystals
cryo-cooling, 641, 661
introduction of anomalous scatterers, 669
packing analysis, 671, 677
soaking, 665
solvent content and solvent channels, 641
cubic equation, 526, 876
cubic system, 648
cumulative dwell-time distribution, 204
cumulative dwell-time histogram, 206
cyan fluorescent protein (CFP), 520
cyclic process, 20, 23, 29
cysteine
attachment of fluorophores, 516
counting of, 489
in disulfide bond formation, 317
in sortase reaction, 518
in thiol and maleimide coupling, 842
redox potential, 134
cystine, 134, 486
cytochromes, 116, 134, 343, 482

D

Dalton's law, 53
damping gas, 810
dative bond, 318
daughter ion, 810
de Broglie wavelength, 635, 734
deconvolution, 544, 813, 903, 904
degeneracy, 92
degree of efficiency (in Carnot cycle), 22
denaturation, 372, 377, *See also* (un)folding
density matching, 627, 635, *See also* contrast variation
density modification, 673
deoxyribonucleic acid, 395, *See also* DNA
deoxyribose, 396, 411
depolarization, 138
depression of freezing point, 58
deshielded, 573
detective quantum efficiency (DQE), 739
detergent, 356
deuteration, 590, *See also* H/D exchange
deuterium, *See also* H/D exchange
in kinetic isotope effects, 192
in mass spectrometry, 814
in neutron crystallography, 674
neutron scattering length, 625
deuterium arc lamp, 480
dialysis, 62, 372, 643
diastereomers, 287, 289
diathermic, 3
dichroic ratio, 494
dielectric constant, 296, 349
difference map, 674
difference Patterson map, 665
difference spectrum, 481, 490, 510, 532, 580
differential equation, 151, 172, 175, 242, 536, 827, 828, 883
differential scanning calorimetry (DSC), 864
in protein folding, 375
oligomer unfolding, 869
two-state unfolding transition, 867
Differential Scanning Fluorimetry (DSF), 378
differentiation, 880
diffraction
of electrons, 682, 746
of neutrons, 682
of X-rays, 640
versus scattering, 620
diffraction grating, 481, 508, 655
diffraction pattern, 620, 659, 661, 691, 736, 904
indexing of, 662
diffusion, 110
lateral, 113, 363
diffusion coefficient, 110, 824
and viscosity, 548
from autocorrelation functions, 705
measurement by fluorescence, 531
measurement by scattering, 620
of biological processes, 621
rotational, 548
translational, 620
diffusion-collision model, 381
dihydroxyacetone phosphate, 72, 162, 288
dipolar coupling
in EPR, 606
in FRET, 550
in NMR, 577
dipole force, 776
dipole moment (definition), 296
dipole strength, 478, 500
dipoles
alignment in α-helices, 330
distance vector, 296
head-to-tail and head-to-head interaction, 298
induced and permanent, 295
partial charges, 296
relative orientation, 296
transient, 303
directed movement, *See* Brownian ratchet
dispersion forces, 303, *See* London dispersion force
displacement titration, 526, 861
dissipation
of charges, 351
of heat, 23
of heat and matter, 28, 34
of matter, 110
dissociation
constant, 73
of acids and bases, 75
distance distribution function, 559, 607, 631
distribution
Boltzmann, 87, 89
distance, 631
dwell time, 204
Gaussian, 712
intensity, 616, 691, 751, 783
disulfide bond, 317
shuffling, 372
disulfide isomerase, 311, 383
DNA, 395
bending and kinking, 421
melting, 491
methylation, 399
positive and negative supercoiling, 416, 785
quadruplex, 413
recognition by proteins, 423
recognition by small molecules, 426
stretching, 772
structures and conformations, 409
dot product, 292, 887
double electron-electron resonance (DEER), 606
double mutant cycle, 78
double-beam photometer, 481
double-mixing stopped flow technique, 791
drift tube, 804
dual-focus fluorescence correlation spectroscopy (2f-FCS), 705
dwell time, 204, 711, 714, 779
dwell-time histogram, 205
dynamic equilibrium, 6, 705
dynamic light scattering, 620

E

Eadie-Hofstee plot, 218
ebullioscope, 60
effect
  bathochromic, 479
  hydrophobic, 349, 356
  hyperchromic, 480
  hypochromic, 480
  hypsochromic, 479
  inner-filter, 532
  kinetic isotope, 191
  Leidenfrost, 745
  Overhauser, 577
  photoelectric, 480
  relativistic, 654
  thermophoretic, 528
  Zeeman, 570
effective concentrations (in catalysis), 282
effective mass, 823, 834
effective substrate concentration, 234
Ehresmann method, 487
Eigenvalues and Eigenvectors, 176, 181, 183, 891
elastic scattering, 534, 615, 624, 741
electrical work, 128
electrochemical cell, 127
electrochemical potential, 126
  and membrane potentials, 136
electrochemical series, 130
electromagnetic spectrum, 472
electromagnetic wave, 471
  polarization plane, 493, 538
electromotive force, 129
electron density, 303, 631, 652
  map, 669, 674
  of biological substances, 627
electron energy loss spectroscopy (EELS), 738
electron gun, 653, 655, 734
electron microscopy (EM), 731
electron multiplier, 811
electron nuclear double resonance (ENDOR), 601
electron paramagnetic resonance (EPR), 599
  and distance measurement, 606
  and dynamics, 603
  and solvent accessibility, 605
electron scattering factor, 741
electron spin, 475, 510, 599
electron spin resonance (ESR), *See* electron paramagnetic resonance (EPR)
electron transfer, 126, 134
electron transport, 134
electron volt (eV), 473
electronegativity, 299
electronic ground state, 475
electronic transition, 473
  absorption, 475
  and solvent, 480, 511
  electron paramagnetic resonance spectroscopy, 599
  fluorescence spectroscopy, 510
  infrared spectroscopy, 504
  Raman spectroscopy, 624
  X-rays, 668
electronically excited state, 475
electrophilic catalysis, 281
electrophoresis, 419, *See also* gel electrophoresis
electrophysiology, 139
electrospray ionization (ESI), 803
electrostatic potential, 126, 423
elementary reaction, 147
elevation of boiling point, 58
elliptically polarized light and elliptic ratio, 497
Ellman's reagent (DTNB), 489
emission spectroscopy, 510
enantiomers, 287, 289
endergonic process, 38
endothermic process, 28
end-to-end distance, 771
energy filter, 738, 742
energy landscape, *See* energy surface
energy minimization, 445, 447, 592, 675
energy potential, 306
  in nucleosides, 406
  of covalent bond, 291
  of non-covalent interactions, 295
  of proteins, 328
energy surface, 290, 328, 339, 373, 406, 437, 447, 769
energy terms, 290, 303, 431, 444, 450, 592, 675
  distance dependence, 290
energy transfer, 550
energy-momentum relationship, 734
enthalpy *H*, 25, 857
  relation to molecular properties, 44
enthalpy-driven reaction, 39, 378
enthalpy-entropy compensation, 40, 344, 349
entropy production, 30, 31
entropy *S*, 28
  and heat, 29
  and order (statistical interpretation), 34
  relation to molecular properties, 44
  temperature dependence, 32
entropy transport, 32
entropy-driven reaction, 39, 378
enzyme
  competitive inhibition, 235
  inhibition at 50% (IC_{50}), 240
  measurement of activity by spectroscopy, 487
  mixed inhibition, 238
  non-competitive inhibition, 236
  non-interacting active sites, 224
  product inhibition, 232, 278, 862
  steady-state kinetics, 213, 861
  time-dependent inhibition, 242
epimer, 289
equation
  Adair, 102
  Arrhenius, 266
  Boltzmann, 35
  cubic, 526, 876
  differential, 151, 172, 175, 242, 536, 827, 883
  Eyring, 268
  Gibbs-Helmholtz, 39
  Henderson-Hasselbalch, 76
  Hill, 228
  Karplus, 575
  Lamm, 828
  Michaelis-Menten, 215
  Nernst, 131
  Perrin, 549
  Porod, 633
  quadratic, 74, 198, 855, 876
  Stern-Volmer, 536
  Stokes-Einstein, 824
  Svedberg, 827
  Thomson, 624
  van't Hoff, 70
equilibrium, *See also* thermal equilibrium
  and Gibbs free energy, 38
  binding and dissociation, 72
  dialysis, 62
  effect of temperature, 72
  protolysis, 75
  titration, 523

equilibrium constant, 69
apparent, 81
determination by thermophoresis, 528
determination by titration, 524
determination by titration, 540
microscopic and macroscopic, 83, 102
equilibrium response (in SPR), 840
ergodicity, 90, 204, 453
error propagation, 218, 897
evanescent field, 697, 838
evaporation, 49, 58
exergonic process, 38
exothermic process, 28
expansion work, 6, 12, 17, 24, 41
expect value (E-value), 815
exponential decay, 151
exponential function, 878
expressed protein ligation, 518, 593, 844
extensive state function, 4
extinction coefficient, 478
and overlap integral, 551
calculation from protein sequence, 485
definition, 479
determination, 485
of biological chromophores, 481
of biological fluorophores, 514
of organic fluorophores, 516
of retinal, 483
Eyring equation, 268

F

F_1-F_o ATPase, 116, 135
FAD, 482
and respiratory chain, 134
as enzyme cofactor, 343, 714
as fluorophore, 515
intramolecular quenching, 535
Faltung, *See* convolution
Faraday constant, 116, 126
fast-exchange regime, 594
fatty acids, 289
ferredoxin, 130, 135
Fick's first law, 110
Fick's second law, 111, 828
fiducial marker, 758
field emission
of electrons, 734
of Ga^+ ions, 746
fingerprint region
IR spectroscopy, 508
NMR spectroscopy, 589
first law
Fick's, 110
Newton's, 826
of Gay-Lussac, 8
of thermodynamics, 7
first-order rate law, 152
half-life, 153
in radioactivity, 153
integrated, 153
first-order reaction, 146
fish
buoyancy, 9
osmoregulation, 63
fission reactor, 635
flash photolysis, 795
flavin adenine dinucleotide, 482, *See also* FAD *and* flavins
flavins, 483, 487, 515
flexizyme, 322, 434
fluid mosaic model, 114, 363
fluorescein, 513, 530, 549, 556
fluorescence, 510
anisotropy, 546
energy transfer, 550
excitation and emission spectra, 512
lifetime, 543
polarization and anisotropy, 537
quenching, 535
time-resolved, 542
fluorescence correlation spectroscopy (FCS), 701
to monitor binding events, 703
fluorescence cross-correlation spectroscopy (FCCS), 707
fluorescence imaging with one nanometer accuracy (FIONA), 712, 724
fluorescence lifetime, 513
fluorescence lifetime imaging (FLIM), 700
fluorescence microscopy, 687
fluorescence recovery after photobleaching (FRAP), 531
fluorescence sensing, 556
fluorescent labeling, 515
fluorimeter, 512
fluorophore
2-aminopurine, 522, 529
attachment to cysteine, 516
cytosine derivatives, 523
extrinsic, 515, 822
FAD and FMN, 515
in membrane, 537, 541
intrinsic, 514
mant, 522
NADH, 515
pyridoxal phosphate, 515
quenching, 535
Y base, 515
folding
of proteins, 369
of RNA, 436
transition, 374
forbidden transitions, 476
force constant, 294, 505, 808, *See also* spring constant
force field, 444
force jump, 779
force measurements, 766
magnetic tweezers, 782
of binding energies, 775
on membrane proteins, 774
optical tweezers, 776
force spectroscopy, 770
force-extension curve, 771
form factor, 619, 625
formalism
Briggs-Haldane, 216
Langmuir-Hinshelwood, 274
Michaelis-Menten, 214, 861
formylglycine generating enzyme, 316, 517
Förster resonance energy transfer (FRET), 550
Förster, T, 550
Fourier coefficients, 664, 665, 674, 900, 902
Fourier series, 900
Fourier shell correlation (FSC), 664, 762
Fourier slice theorem, 758
Fourier transformation, 508, 584, 608, 758, 901
of Coulomb potential density, 741
of point spread function, 751
of structure factor, 660, 664
Fourier transform-IR (FT-IR), 508
Fourier transform-NMR (FT-NMR), 585
Franck-Condon principle, 476
Fraunhofer diffraction pattern, 691
free electron laser (FEL), 657, 662, 680
free energy G, 37
change and equilibrium constant, 68

relation to molecular properties, 44
temperature dependence, 41
free induction decay (FID), 580
free *R*-factor, 676, 678
freely jointed chain (FJC) model, 771
freezing point depression, 58
frequency, 472
radial, 762
spatial, 748
frequency-modulated mode (in AFM), 767
FRET efficiency, 551, 552, 558, 561, 721
friction, 822
coefficient, 548, 620, 824
Friedel's law, 667
function
partition, 87, 90, 92, 870
softmax, 455
functional magnetic resonance imaging (fMRI), 598
futile cycle, 23, 120

G

galactose, 289
Gaussian chain (GC) model, 771
Gaussian distribution, 895
convolution, 903
in mass spectrometry, 814
in optical trapping, 776
in ultracentrifugation, 836
of fluorescence intensity, 544, 692, 712
three-dimensional, 559
gel electrophoresis, 419, 585, 815
general gas constant, 4
genetic code, 320, 404
g-factor (in EPR), 599, 603
G-factor (in fluorescence), 539
Gibbs free energy, 38, *See also* free energy *G*
Gibbs phenomenon, 901
Gibbs-Helmholtz equation, 39
Gill & von Hippel method, 485
global distance test (GDT), 461
globin fold, 342
glucose, 289, 362, 710
glutamate, 281, 317, 796
glutaredoxin, 134, 317
glutathione, 130
as redox buffer, 134
glutathione S-transferase (GST), 160, 323, 345
glyceraldehyde-3-phosphate, 72, 162, 288
glycerolipids, 354
glycogen, 315
glycolysis, 168
glycosidic bond, 396
glycosylation, 314
goniometer, 661
G-quadruplex, 413
green fluorescent protein (GFP), 337, 464, 520
group-specific inhibitor, 249
GSH reductase, 134
Guinier equation and Guinier plot, 630
gyrase, 420, 421
and magnetic tweezers, 783
and single-molecule FRET, 720
inhibited by novobiocin, 236

H

H/D exchange
in neutron crystallography, 674
in NMR, 594, 596
to follow complex formation, 815
to follow protein folding, 387
Hahn echo, 606
half-cell, 127
half-life, 153
halogen bonds, 302
hangover, 136, 172
Harker construction, 666
harmonic oscillator, 291, 504
heat, 6, 20
and entropy *S*, 29
latent, 33
heat capacity, 17, 26, 864
heat denaturation, 377
heat shock proteins, 385
helical junction
in DNA, 414
in RNA, 429
studied by FRET, 720
helix dipole, 330, 351
effect on pK_A, 385
in DNA binding, 426
in protein folding, 383
helix wheel, 332
helix-turn-helix motif, 423
hemoglobin, 229, 343, 346, 482
Henderson limit, 647
Henderson-Hasselbalch equation, 76
heptad repeat, 331
Hess's law, 27, 867
hexagonal system, 648
hexokinase, 352, 362
hidden Markov modeling, 714
highly-inclined and laminated optical sheet microscopy (HILO), 698
high-pass filter, 750
Hill coefficient, 231
Hill equation, 228
Hofmeister series, 370
Holliday junction, 414
studied by three-color FRET, 720
homolog, 325
homologous recombination, 414
homology modeling, 456
homonuclear and heteronuclear coupling, 587
Hoogsteen edge, 399, 411, 412, 413, 423, 436
Hooke's law, 291, 504, 767
horror vacui, 370
hybrid methods, *445*
hydrodynamic bead modeling, 825
hydrodynamic radius, 629
and DNA topoisomers, 419
hydrogen
in H/D exchange, 387
neutron scattering length, 625
hydrogen bonds, 299
between nucleobases, 397
energy, 301
importance for molecular recognition, 351
intramolecular, 302, 353
networks and cooperativity, 301
partial covalent character, 300
hydrogen electrode, 129
hydrophobic collapse model, 381
hydrophobic effect, 349, 356
hydroxylation, 316
hyperchromic effect, 480, 491
hyperfine coupling, 601
hyperpolarization, 138
hypertonic, 63
hypochromic effect, 480, 487, 491
hypothesis
chemiosmotic, 116

RNA-world, 434
wobble, 401
hypotonic, 63
hypsochromic effect, 479

I

ideal gas, 8
ideal gas law, 4
ideal solution, 56
image contrast, 749
immunoglobulin, 318, 342
indole (in Trp), 299, 310
induced dipole, 295, 303, 616
induced fit, 352
inelastic scattering, 615, 622, 668, 737, 742
infrared spectroscopy, 504
inhibitor
allosteric, 237, 353
competitive, 235, 240
group-specific, 249
non-competitive, 236
slow-onset, 242
suicide, 249
inner-filter effect, 497, 532
insertion devices (wiggler and undulator), 656
in-source decay (ISD), 810
instrument response, 544, 751, 861
integral membrane protein, 364, 483
integrated rate law, 150, 203
integration, 880
integrative methods, 445
intensity distribution, 691, 751, 783
intensive state function, 4
interactions, 290
between induced dipoles, 303
between ions and dipoles, 296, 299
between permanent dipoles, 298
covalent, 290
electrostatic, 290
energy and distance dependence, 295
intramolecular, 290
ionic, short-range, and long-range, 295
non-covalent, 290, 295
polar *versus* non-polar, 295
intercalation, 419, 426
interference pattern, 620, 627, 778, 783
interferogram, 508
intermediate-exchange regime, 594
intermolecular
coupling of spins, 607
energy, 44
fluorescence quenching, 522
hydrogen bonds, 302
triplex and quadruplex, 413
vectors, 665
internal conversion (IC), 510
internal energy U, 6
temperature dependence, 24
inter-system crossing (ISC), 510
intramolecular, *See also* intermolecular
catalysis, 282
Förster resonance energy transfer, 555
hydrogen bonds, 302, 353
quenching of fluorescence, 535
intrinsically disordered proteins, 374
inversion symmetry, 664
inverted repeats, 414
ion channel, 139, 483, 796
ion condensation, 426
ion detection, 811
ion fragmentation, 809
ion mobility, 816
ion transport, 118
ion trap, 805, 816
ionic interactions, 296
ionization energy, 291, 734
ionization enthalpy of buffers, 858
IR region, 472
irreversible isothermal expansion, 11
isobaric conditions, 8, 25
isobaric tags for relative and absolute quantification (iTRAQ), 818
isochoric conditions, 8, 16
isoelectric point (pI), 324
isolated system, 3, 30, 37
isomers, 287
anomers, 289
classification, 288
conformers, 290
diastereomers, 289
tautomers, 288, 396
isomorphous replacement, 665
isopycnic centrifugation, 834
isotherm, 8
isothermal expansion, 11, 12
isothermal titration calorimetry (ITC), 853
isotope abundance, 572
isotope exchange, 814
isotope labeling, 590, 593, 819, 836
isotopolog, 818, 877

J

Jablonski diagram, 475, 477, 510, 511, 550
J-coupling, 575, 592
jellyfish (*Aequorea victoria*), 520
Joule's law, 17

K

k turn, 429
Karplus equation, 575
kinase, 237, 315, 325, 327, 352, 362, 488
kinetic competition, 165, 167
kinetic control, 270
kinetic crystallography, 662, 679
kinetic energy, 6
and activation energy, 269
of ions, 804
of relativistic electrons, 735
temperature dependence, 450
kinetic isothermal titration calorimetry (kinITC), 863
kinetic isotope effect, 191, 268, 506
kinetic partitioning, 168, 190, 261
kinetic trap
in protein folding, 373, 380
in RNA folding, 437
Kirchhoff's law, 26, 868
Kirkwood approximation, 826
kissing loop, 429, 646
Kratky plot, 633
K-shell, 653

L

l'Hôpital's rule, 886
lactate dehydrogenase (LDH), 345, 488
Lambert-Beer law, 478, 834
determination of extinction coefficients, 485
deviations from, 492
Lamm equation, 828
Langmuir-Hinshelwood formalism, 274

Larmor frequency, 572, 597
laser flash photolysis, 795
latent heat, 33
lateral diffusion, 113, 363
lattice plane, 658
Laue diffraction, 662, 679
law
  0th, of thermodynamics, 6
  1st, of thermodynamics, 7
  2nd, of thermodynamics, 22, 28, 30
  3rd, of thermodynamics, 33
  Avogadro's, 8
  Boyle-Mariotte, 8
  Bragg's, 658
  Charles', 8
  Coulomb's, 296
  Dalton's, 53
  Fick's, 1st, 110
  Fick's, 2nd, 111, 828
  Friedel's, 667
  Gay-Lussac, 1st and 2nd, 8
  Hess's, 27, 867
  Hooke's, 291, 504, 767
  ideal gas, 4
  Joule's, 17
  Kirchhoff's, 26, 868
  Lambert-Beer, 478, 492, 834
  Newton's, 1st, 826
  Raoult's, 53
  rate, 145
  Snellius', 657, 697
  Stokes', 824
leaflet (of membrane), 330, 361
leakage, 120, 139
leap-frog algorithm, 452
left-handed circularly polarized light (LHCPL), 496
Leidenfrost effect, 745
Lennard-Jones potential, 303
lens, 688
  electromagnetic, 736, 806
leucine zipper, 331, 424
Levinthal paradox, 373, 460
lifetime, 153
  and quantum yield, 513
  of fluorescence, 543
Lifson-Roig model (of protein folding), 383
ligand-gated channel, 118, 139, 483
light scattering, 615
light-sheet microscopy, 698
limiting anisotropy, 548
line shape analysis, 576, 594, 596
line width, 576, 594, 600, 903
linear dichroism (LD), 493
linearly polarized light, 493
line-broadening
  in EPR, 606
  in NMR, 593, 597
Lineweaver-Burk plot, 218
linked functions, 76
  and apparent equilibrium constants, 81
linking number, 417
lipid, 355
  bilayer (membrane), 289, 354, 361
  fluorescent labeling, 523
  monolayer, 356
lipid rafts, 114, 363
lipidic cubic phase (LCP), 646
liposome, 362
local magnetic field, 572
local minimum, 119, 373
local resolution (in EM maps), 763
local symmetry, 651, *See* non-crystallographic symmetry
lock-and-key model, 352
locked nucleic acid (LNA), 411
logarithm, 878
London dispersion force, 303
lone pair, 296, 299, 318
longitudinal relaxation time, 581, 600
long-pass filter, 694
long-range interactions, 295
Lorentz force, 656, 681, 736, 808
low complexity region (LCR), 323, 343, 463
low-pass filter, 758
L-shell, 653
lung, 9

M

machine learning, 454
Maclaurin series, 899
macrostate, 34
magic angle spinning (MAS), 598
magnetic resonance, 570
magnetic resonance imaging (MRI), 598
magnetic tweezers, 782
magnetogyric ratio, 571
major groove, 411, 428
  binding of proteins, 423
maleimide coupling, 843
mannose, 289
mass analyzer, 804, 805, 808
mass photometry, 618
mass spectrometry (MS), 801
mass transport limitation, 841
matrix algebra, 175, 888
matrix-assisted laser desorption ionization (MALDI), 802
Matthews parameter, 651
maximum likelihood (ML), 676, 760
maximum velocity, 214
mean square displacement, 113, 677
melanin, 483
melatonin, 259
melting temperature, 865
  and base stacking, 408
  of protein, 375
membrane, 354
  as solvent for proteins, 354
  asymmetry, 363
  biological roles, 116, 354
  dialysis, 62
  fluidity, 289
  fusion, 362
  potential, 136
  rafts, 114, 363
  semipermeable, 60
  vesicle, 356
membrane proteins, 354
  crystallization of, 645
  examples, 365, 683
  folding of, 381, 774
  lateral diffusion, 113, 713
  properties and structures, 364
  solubilization, 367
  studied by AFM, 774
  studied by EPR, 605, 608
  studied by fluorescence, 537
  studied by mass spectroscopy, 816
  studied by Raman spectroscopy, 624
  studied by SANS, 636
  studied by solid-state NMR, 598
  studied by SPR, 846
Meselson-Stahl experiment, 836

mesomers, 310
messenger RNA (mRNA), 308, 320, 395
metabolic labeling, 818
metal chelation, 317, 319, 530, 549, 796
metal ions
  binding to nucleic acids, 413, 426
  binding to proteins, 318, 325
  detection by electron paramagnetic resonance, 601
  in electrochemistry, 127
  in electrophilic catalysis, 281
  in macromolecular crystallization, 646
  interaction with aromatic systems, 299
metal shadowing, 744
metastable compound, 270
methanethiol sulfonate spin label (MTSSL), 601
micellar extraction, 367
micelles, 356
Michaelis-Menten formalism, 214, 861
Michelson interferometer, 508
microchannel plate detector (MCP), 812
microscale thermophoresis, 528
microscope, 687
  magnification, 690
microscopy
  atomic force, 766
  confocal, 695, 723
  electron, 731
  fluorescence, 687
  light-sheet, 698
  near-field, 693
  Raman, 623
  scanning, 766
  single-molecule, 708
  spinning disc confocal, 696
  super-resolution, 723
  total internal reflection, 697
  wide-field, 693
microstate, 34
microwave region, 472
Miller indices, 658
MinFlux, 712
minor groove, 411, 428
  binding of proteins, 425
  ligand binding, 426
missing wedge, 756
mitochondria, 116, 724
mixed inhibition, 238
model
  concerted, of cooperativity, 229
  conformational selection, 352
  fluid mosaic, 114
  in protein folding, 381
  induced fit, 352
  Koshland-Nemethy-Filmer (KNF), 106
  lipid raft, 114
  lock-and-key, 352
  Monod-Wyman-Changeux (MWC), 105
  sequential, of cooperativity, 229
  to describe elastic polymers, 771
  two-state, 374
  worm-like chain (WLC) model, 771
model bias
  in AlphaFold2, 460, 464
  in cryo-EM, 760
  in cystallography, 672
molar concentration, 55
molar ellipticity, 499
molar extinction coefficient, 479
molar property, 5, 47
molar thermal energy, 474
mole fraction, 53
  effective, 57
molecular dynamics, 443, 449
molecular form factor, 626
molecular interfaces
  energy involved in, 349
  properties of, 347
molecular mass, 801
  from light scattering, 618
  from mass spectrometry, 813
  from ultracentrifugation, 827, 831, 836
  from X-ray scattering, 633
molecular modeling, 443
  application to macromolecules, 456
  coarse-grained modeling, 448
  homology modeling, 456
  in SAXS, 634
  simulated annealing, 447
  using FRET restraints, 721
molecular motor, 124
molecular polarizability, 299, *See also* polarizability
molecular recognition, 302, 351, 423, 426
molecular replacement, 670
molecularity, 147
molecule detection function (MDF), 692
molten globule, 382
momentum transfer, 628
monoclinic system, 648
Morse potential, 291
mosaicity, 662
Mott-Bethe formula, 741
M-shell, 653
multi-angle light scattering (MALS), 620
multiple sequence alignment (MSA), 327, 340, 458
myoglobin, 343, 482

N

NADH, 482
  absorption properties, 488
  and respiratory chain, 134
  as enzyme cofactor, 343
  as fluorophore, 515
  intramolecular quenching, 535
  reduction to NAD^+, 126, 130
NADPH/$NADP^+$, 116, 135
nanodisc, 368
native mass spectrometry, 815
natural line width, 596
near-field microscopy, 693
negative cooperativity, 228
negative staining, 745
Nernst equation, 131
neural network, 454
neutron crystallography, 674
neutron match point, 627, 636
neutrons, 635
Newtonian equations, 451
Newton-Raphson algorithm, 446
nicotinamide adenine dinucleotide, 482, *See also* NADH
non-competitive inhibition, 236
non-contact mode (in AFM), 767
non-crystallographic symmetry (NCS), 651, 673
non-ideal solution, 56, 69, 626
non-polar interactions, 295
normal distribution, *See* Gaussian distribution
normalized discrepancy function (in SAXS), 634
normalized state variable, 5, 47
nuclear density, 627, 682
nuclear magnetic resonance (NMR), 570
  and dynamics, 594
  and imaging, 598

solid state, 597
nuclear magnetogyric ratio, 571
nuclear Overhauser effect (NOE), 577, 592
nuclear Overhauser enhancement spectroscopy (NOESY), 588
nuclear spins, 570, 580
nuclear transition, 571
nucleation model, 381
nucleation-condensation model, 381
nucleic acids, 395
absorption, 482
circular dichroism of, 501
composition of, 396
concentration determination, 487
fluorescent labeling, 515
primary structure, 403
spin labeling, 601
surface immobilization, 708
nucleobases, 396
non-standard, in DNA, 399
non-standard, in RNA, 400
nucleophilic catalysis, 281
nucleosides and nucleotides, 396
syn/*anti* conformation, 401
nucleosome, 421
numerical aperture (NA), 688
Nyquist frequency and limit, 740

O

observed rate constant, 194
ocular, 690
oligomer, 345
one-step binding, 194
rate constant and amplitude, 197
open system, 3, 46, 71
operator, 110, 446, 904
opsin, 483
optical activity, 496
optical diffraction limit, 691
optical filter, 694
optical rotation dispersion (ORD), 497
optical tweezers, 776
dumbbell configuration, 780
force-clamp and extension-clamp modes, 779
optogenetics, 484
orbitrap, 808
orientation factor (in FRET), 552
ortholog, 325, 403
oscillator, 291, *See* harmonic oscillator
osmoregulation, 63
osmosis and osmotic pressure, 60
osteogenesis imperfecta, 388
overall order of a reaction, 146
Overhauser effect, *See* nuclear Overhauser effect
overlap integral, 551
oxidation, 126
oxidative phosphorylation, 116
oxygen
as collisional quencher, 535
as electron acceptor, 126, 134
as triplet state quencher, 710
binding to hemoglobin, 229, 343
in hydrogen bonds, 300
paramagnetism, 605
scavengers, 710
singlet state, 603

P

Pake pattern, 608
palindrome, 414
parabolic potential energy, 291
parallel reactions, 166
parallel β-sheet, 334
paralog, 325, 403
paramagnetic ions, 573, 605
paramagnetic relaxation enhancement, 597
parent ion (in MS), 810
partial charges of atoms, 296
partial derivative, 17, 47, 881
partial fractions, 158, 885
partial molar property, 51
partial pressure, 53
partial specific volume, 823
particle picking, 759
partition function, 87, 90, 870
for degenerate states, 92
in all-or-none model, 96
on ensemble and molecular levels, 93
Pascal's triangle, 35, 226, 876
passband, 754
passive transport, 118
patch tracking, 758
patch-clamp method, 139
Patterson function, 664, 671
peptide bond and peptide group, 311
peptide groups
alignment of dipoles in α-helices, 330
and metal binding, 318
cis/*trans* isomers, 312
peptide nucleic acid (PNA), 411
peripheral membrane protein, 364
permanent dipole, 295
permittivity, 296, *See also* dielectric constant
Perrin equation, 549
persistence length, 772
pH, 76
and biochemical standard state, 44
and enzyme kinetics, 220
measurement, 133
phase contrast, 749
phase diagram, 48, 642
phase plate, 750
phase problem, 640, 660
phase separation, 350
phase transition, 33
in lipid bilayers and nucleic acids, 50
measurement by DSC, 864
phosphatase, 315
phosphatidylserine, 363
phosphodiester, 281, 397, 403
phospholipids, 355
and diffusion, 363, 531, 621
phosphonate (as transition state analog), 249
phosphorescence, 510, 662
phosphorylation, 315, 325
of serine, threonine, and tyrosine, 315, 797
oxidative, 116
photoactivated localization microscopy (PALM), *725*
photobleaching, 521, 531, 709, 722
photodiode, 687, 695, 708
photoelectron, 641, 700
photometer, 480
photon counting histogram (PCH), 706
photoselection, 493, 538
photosynthesis (photosystems I and II), 116, 135
pi (π)-bond, 289
ping-pong mechanism, 260
Planck constant, 268
plasmid, 416, 419, 836
plasmolysis, 60, 63
plasmon resonance, 837

plastoquinone, 116, 135
pleated sheet, 336
P-loop (for nucleotide binding), 351
plot
  chevron, 386
  Eadie-Hofstee and Lineweaver-Burke, 218
  Guinier, 630
  Kratky, 633
  Ramachandran, 328, 339, 406, 678
  Stern-Volmer, 536
  van't Hoff, 70
point-spread function (PSF), 692, 751
polarizability, 299, 303, 476, 616, 618, 622, 624, 776
  and refractive index, 618
polarization factor, 618
polarized light, 493
polarizer, 500, 539
polydispersity, 829
polynucleotide, 395
polypeptide, 308
poly-Pro helix, 332
population
  inversion, 581
  of states, 474, 510, 577
Porod equation and Porod invariant, 633
porphyrin, 483
positive cooperativity, 228
post-source decay (PSD), 810
post-translational modification, 313
potential
  chemical, 71
  electrochemical, 126
  electrostatic, 423
  Lennard-Jones, 303
  Morse, 291
  of membranes, 136
  of standard electrode, 130
potential density, 756
potential function
  6–12 potential, *See also* Lennard-Jones
  and bond angle, 292
  and bond length, 291
  and force, 832
  Coulomb, 296
  energy terms, 290
  for *cis/trans* isomers and torsion angles, 294
  for dipolar interactions, 297, 303
  in molecular modeling, 444
  Lennard-Jones, 303
  London, 303
  saw-tooth, 120
  van der Waals, 303
potential map, 741, 763
power series, 899
power spectrum, 751
precession, 572
predicted aligned error (PAE), 461
predicted local distance difference test (pLDDT), 461
pre-exponential factor, 268
pressure, 4
pressure dependence
  of chemical potential μ, 51
  of equilibrium constant, 72
  of Gibbs energy, 42
pressure jump, 387, 797
primary inner-filter effect, 532
primary kinetic isotope effect, 191
principle
  Archimedes', 823
  ergodic, 204, 453
  Franck-Condon, 476
  Le Chatelier's, 71, 216, 387, 769
principle component analysis (PCA), 760
principle of vertical transitions, *See* Franck-Condon principle
probability
  and equilibrium constant, 89
  and free energy, 88
  of states, 87
  on ensemble and molecular levels, 90
probability density, 711, 897
product inhibition, 232, 278, 862
product rule, 880
projection slice theorem, 758
prolyl isomerase, 383, 792
protease, 83, 249, 261, 281, 311, 385, 388, 674, 815
protein disulfide isomerase (PDI), 311, 383
protein domains, 341, 344
protein folding, 369
  and force methods, 769, 778
  diseases, 388
  driving forces, 369
  energy landscapes, 373
  intermediates, 381, 382
  kinetics, 373, 383, 386
  monitored by spectroscopy, 375, 490
  of oligomers, 382
  of RNase, 372
  pathways, 380
  rate-limiting steps, 383
  two-state model, 374
protein stability curve, 377
protein-induced fluorescence enhancement (PIFE), 715
protein-ligand interactions, 351
protein-protein interactions, 344, 347
proteins, 308
  circular dichroism of, 501
  concentration determination, 485, 487
  crystallization of, 641
  fluorescent labeling, 515
  stabilization by ligand binding, 379
  surface immobilization, 708, 842
protolysis, 75, 857
proton gradient, 116
proton pump, 135, 483
proton transfer, 281
proton-motive force (PMF), 116
pseudo-atom, 437, 448
pseudo-first-order reaction, 195
pseudo-knot, 429
pseudo-symmetry, 346
pulsed electron-electron double resonance (PELDOR), 606
pulsed interleaved excitation (PIE), 722
pyridoxal phosphate, 281, 352, 501, 515
pyrrolysine, 320
pyruvate, 168
  carboxylase, 261
  kinase (PK), 237, 488

Q

Q tag, 517
quadratic equation, 74, 198, 855, 876
quadruplex (in DNA), 413
quadrupole, 805, 816
quantitative mass spectrometry, 818
quantization of energy, 506
quantum dot, 521
quantum number
  spin, 572
  vibrational, 506
quantum yield, 513
  of organic fluorophores, 513, 516

of protein residues, 514
quaternary structure, 327, 344, 382
quench flow technique, 794
quenching
by oxygen and polarizable compounds, 535
collisional and static, 535
in FRET, 553
of chemical reactions, 794
of spin labels (in EPR), 605
of triplet state, 710
quotient rule, 880

R

R/S convention, 289, 308
radiation damage, *See also* photobleaching
by electrons, 742
by neutrons, 635
by UV light, 483
by X-rays, 632, 641, 647, 669
in fluorescence, 709
radiative decay, 542, 550
radical mechanism, 147, 191
radio wave region, 472
radiocarbon method, 153
radioisotopes, 154
radius of gyration and radius of hydration, 629
Ramachandran plot, 328, 339, 406, 678
Raman scattering, 622, 624
Raman spectroscopy, 622
random coil, 383, 502, 574, 629
Raoult's law, 53
rapid equilibrium approach, 214
Ras protein, 351, 509
rate constant, 146
apparent, 164, 166, 386
net, 261
observed, 194
temperature dependence of, 266
rate law, 145, 203
for consecutive reactions, 168
for first-order reactions, 150
for HBr formation, 147
for heterogeneous catalysis, 276
for HI formation, 147
for parallel reaction, 166
for reversible reactions, 162
for second-order reactions, 155
for time-dependent inhibition, 248
in steady-state kinetics, 213
integrated, 150, 151, 152, 203, 248
integrated with matrix algebra, 175
rate-limiting steps, 189, 214, 277, 383, 644
Rayleigh criterion, 693
Rayleigh ratio, 618
Rayleigh scattering, 618, 624
reaction
elementary, 146
enthalpy- and entropy-driven, 39, 378
first-order, 146
molecularity, 147
networks, 172
pseudo-first-order, 195
second-order, 146
steady-state, 213
reaction center (in photosynthesis), 135
reaction coordinate, 273
reaction order, 146
reaction scheme, 255
complex variants, 261
for acid-base catalysis, 278
ordered or sequential binding, 257
ping-pong, 260
random binding, 255
reaction types, 162
parallel, 166
redox, 126
reversible, 162
sequential or consecutive, 168
reaction velocity, 145, 150, 862
reactive oxygen species (ROS), 134, 710
real-time NMR, 596
reciprocal distance, 748
reciprocal lattice, 659
reciprocal space, 508, 628, 659, 746, 757, 758, 762
recognition, 302, *See also* molecular recognition
redox pair, 129
redox reactions, 126
in biology, 134
redox-titration, 603
reduced linear dichroism (RD), 494
reduced mass, 505
reduction, 126
reference state, 39, 43, 51, 56, 351
reflection
of neutrons, 635
of X-rays, 654, 657
symmetry-related, 662
total internal, 697, 838
unique, 662
reflectron, 804
refractive index, 551, 618, 688, 697, 837
relative centrifugal force (RCF), 823
relative dielectric constant, 296
relativistic effect, 654
relaxation kinetics, 797
relaxation time, 576, 600
relaxed state, 421
resolution
diffraction-limited, 691, 737
in cryo-EM, 762
in scanning transmission electron microscopy, 732
in X-ray crystallography, 658, 663, 673
of a microscope, 692
of magnetic and optical tweezers, 770
of mass spectra, 804, 812
resonance energy transfer, 550
resonance frequency, 594
resonance Raman scattering, 624
respiratory chain, 134
resting potential, 139
retinal, 289, 483, 624, 795
reverse osmosis, 63
reverse turn, 337
reversible isothermal expansion, 12
reversible reaction, 162
reversible saturable optically linear fluorescence transitions (RESOLFT), 724
R-factor, 676, 678
rhodopsin, 483
rhombohedral system, 648
ribonuclease, *See* RNase
ribonucleic acid, 395, *See also* RNA
ribose, 281, 396, 401, 428, 522, 590
ribosome, 437, 637, 817
riboswitch, 434
ribozyme, 234, 276, 277, 395, 403, 434
right-handed circularly polarized light (RHCPL), 496
rigid-body refinement, 677, 763
ring current, 573
RNA, 395, 428
catalytic (ribozymes), 434

folding of, 436, 770
secondary structure prediction, 431
tertiary structure, 431
RNA-ligand interactions, 434
RNA-protein interactions, 437, 816
RNase, 281, 352, 372, 384, 792, 793
Röntgen, W., 653
root mean square distance (rmsd), 875
RoseTTAFold, 456
Rossmann fold, 342
rotamer, 313
rotating anode, 654
rotation function, 671
rotation matrix, 650, 670
rotational correlation time, 546, 593, 603
rotational strength (in circular dichroism), 500
rule of thumb
for charges of ionized proteins, 803
for design of titration calorimetry experiments, 860
for free *R*-factor, 677
for reaction velocities, 266
for rotational correlation times, 548
rupture force, 775

S

salt bridge, 349
sampling frequency, 740
sarcin-ricin loop, 429
saturation (of enzymes), 73, 214
saturation response (in SPR), 840
saturation transfer difference NMR (STD-NMR), 579
saw-tooth-shaped energy potential, 120
scalar coupling, 575
scalar product, 599, 887
scalar triple product, 888
scanning confocal microscopy (SCM), 695
scanning electron microscopy (SEM), 731, 744
scanning tunneling microscopy (STM), 766
scattering, 615
and molecular mass, 618, 633
and molecular shape, 634
and optical tweezers, 776
distortion of absorption spectra, 492
distortion of fluorescence spectra, 534
elastic *versus* inelastic, 615, 731, 741
intensity distribution, 616, 629
of electrons, 741
of electrons, 737, 741
of visible light, 615
of X-rays and neutrons, 624
Raman, 622, 624
Rayleigh, 618, 624
resonance Raman, 624
scattering density, 626, 636
scattering length, *See* Thomson scattering length
of various isotopes, 625
scattering length density, *See* scattering density
scattering pattern, *See* interference pattern
scattering vector, 628, 633
Scherzer focus, 754
scrambled RNase, 372
screw axis, 650
scuba diving, 9
sea anemone (*Discosoma striata*), 520
sealed tube, 654
second law
Fick's, 111, 828
of Gay-Lussac, 8
of thermodynamics, 22, 28, 30
secondary inner-filter effect, 532
secondary ion mass spectrometry (SIMS), 817
secondary kinetic isotope effect, 191
secondary structure, 327
detected by CD spectroscopy, 501
detected by IR spectroscopy, 509
detected by NMR, 574
formation during protein folding, 381, 383
helical (in proteins), 330
of DNA, 409
of RNA, 429
prediction (of proteins), 340
prediction (of RNA), 431
strands (in proteins), 334
turns (in proteins), 337
second-order rate law, 155
half-life, 156
integrated, 155
second-order reaction, 146
sector-shaped cell, 821
sedimentation, 820
opposed by diffusion, 831
sedimentation coefficient, 820, 827
concentration dependence, 828
determination, 827
sedimentation equilibrium, 820, 831
sedimentation velocity, 820, 826
selection rule, 476
selenocysteine, 320
self-amplified spontaneous emission (SASE), 681
semipermeable membrane, 60
sensorgram (in SPR), 839
separation of variables, 196, 478, 827
sequence alignment, 325, 672
sequence analysis, 322, 809
sequential model of cooperativity, 229, 276
serial crystallography, 680
serine
glycosylation, 314
in proteases, 83
phosphorylation, 315, 797
shadowing (in electron microscopy), 744
Shannon sampling theorem and limit, 740
shape complementarity, 348
shielding constant (in NMR), 572
short-pass filter, 694
short-range interactions, 295
sigma (σ)-bond, 289
sigma (σ)-hole (in halogen bonds), 302
sign convention
for energy, 3
for heat and work, 7
signal-to-noise ratio, 739, 749, 797, 859, 895
simulated annealing, 447, 592, 634, 677
sine rule, 877
single-injection method, 859
single-mixing stopped flow technique, 791
single-molecule
advantages, 711
fluorescence microscopy, 708
FRET, 561, 718
kinetics, 203, 712, 714
pull-down (SiMPull), 717
singlet state, 475
size exclusion chromatography-multi-angle laser light scattering (SEC-MALLS), 618
slow exchange regime, 594
slow-onset inhibition, 242
small angle scattering (SAS), 624
of neutrons (SANS), 635
of X-rays (SAXS), 632
SNAP tag, 517

Snellius' law, 657, 697
sodium dodecylsulfate (SDS), 370
softmax function, 455
solid-state NMR, 597
solubility curve, 644
solute and solvent, 55
solvation
  and absorption, 480
  and fluorescence, 533
  and polar interactions, 296, 301, 366
  in molecular modeling, 450
  of fluorophors, 511
  of non-polar surfaces, 865
  of unfolded state, 370
solvent matching, 636, *See also* contrast variation
solvent relaxation, 511
solvent-accessible surface area, 348
sortase reaction, 518
space groups, 647
  tabulation of, 650
spallation source, 635
spatial coherence, 734
spatial frequency, 628, 748
specific heat capacity, 5
specific linking difference, 420
spectrometer, 480
spectroscopy
  absorption, 475, 493
  emission, 510
  fluorescence correlation (FCS), 701
  infrared (IR), 504
  linear and circular dichroism (LD and CD), 493
  nuclear magnetic resonance (NMR), 570
  principles of, 473
  Raman, 622
spectrum (electromagnetic), 472
spherical aberration, 737
spin, 510, 570, 599
spin label, 601
spin quantum number, 572
spin system, 588
spin trap, 601
spine of hydration (in DNA), 411
spin-lattice relaxation time, *See* longitudinal relaxation time
spinning disc microscopy, 696
spin-spin relaxation time, *See* transverse relaxation time
spliceosome, 437
splicing
  in proteins, 518
  in RNA, 437
splitting
  of EPR resonances, 601
  of NMR resonances, 575
  of spin states (Zeeman effect), 570
spontaneous process, 28
  and chemical potential, 47
  in diffusion, 110
  inhibition by membranes, 118
  kinetic control of, 270
spring constant, 292, 808, *See also* force constant
stable isotope labeling by amino acids in cell culture (SILAC), 818
stacking
  and fluorescence quenching, 515, 522, 549
  and hypochromism, 491
  and intercalation, 427
  co-axial (of helices), 414, 432, 720
  in RNA, 428
  of nucleobases, 405
staining
  in electron microscopy, 745
  of DNA, 427
standard chemical potential, 51
standard conditions, 43
standard deviation, 662, 836, 895
standard electrode potential, 130
standard electromotive force, 129
standard free energy of formation, 39
standard hydrogen electrode, 129
state function, 4
  extensive and intensive, 4
  independence of pathway, 16
  relation to molecular properties, 44
  under biochemical standard conditions, 44
state variable, 4, 8, 768
  normalized, 5, 47
  partial molar, 52
static light scattering, 615
static quenching, 535
statistical thermodynamics
  and binding equilibria, 100
  and structural changes, 95
statistical weights, 88, 90, 96
steady-state approximation, 216, 278
steady-state kinetics, 213
steepest descent algorithm, 446, 455
step gradient, 820, 835
stereocenter, 289
Stern-Volmer equation, 536
Stern-Volmer plot, 536
stimulated emission depletion (STED), 723
stochastic optical reconstruction microscopy (STORM), 725
stoichiometry, 132
  from EPR, 608
  from fluorescence titration, 525
  from ITC, 855
  from mass spectrometry, 818
  from single-molecule experiments, 717, 722
  in rate laws, 146
  in surface plasmon resonance, 845
  of protein complexes, 344
Stokes band (Raman scattering), 622
Stokes shift, 511, 514
Stokes' Law and Stokes-Einstein equation, 824
stopped flow technique, 791
streak camera, 700
stretch-release curve, 774
structural motif, 327, *See also* super-secondary structure
structure calculation (in NMR), 592
structure factor, 620, 660
structure-based drug design, 353, 652, 665
substrate cycle, 24
substructure, 665
sucrose (gradients in ultracentrifugation), 835
sugar pucker, 401
suicide inhibitor, 249
sulfatases, 316
supercoiling (in DNA), 416
superhelical density, 421
superhelix
  in DNA topoisomers, 418
  in nucleosomes, 421
  in proteins, 331
super-resolution microscopy, 723
super-secondary structure, 327, 341
surface complementarity, 348
surface entropy reduction, 645
surface immobilization, 708, 842
surface plasmon resonance (SPR), 837
  mass transport limitation, 841
  principle and information content, 839
  specificity and stoichiometry of binding, 845
surface-enhanced Raman spectroscopy (SERS), 624

surfactant, 356
surroundings, 3, 37
Svedberg equation, 827
Svedberg, T., 827
symmetry
  and bond stretching, 507
  and crystal lattice, 648
  and space groups, 647, 650
  non-crystallographic, 651
  of oligomers, 345
  operator, 649
symport, 118
synchrotron, 654
synonymous codon, 320
system
  citeria for spontaneous processes, 48
  crystal, 648
  open, closed, and isolated, 3
  two-state (in protein folding), 374
  two-state (in spectroscopy), 473

T

tandem-MS (MS/MS), 808, 810
tapping mode (in AFM), 767
tautomers, 287, 288, 396
Taylor series, 447, 451, 899
temperature dependence
  of absorption, 490
  of chemical potential μ, 51
  of enthalpy *H*, 26, 857
  of entropy *S*, 32, 376
  of equilibrium constants, 70
  of free energy *G*, 41
  of heat capacity C_p, 868
  of internal energy *U*, 17, 24
  of kinetic energy, 450
  of rate constants, 266
temperature jump, 387, 797
temporal coherence, 734
tertiary structure, 327, 341
  formation during protein folding, 381
  of DNA, 414
  of RNA, 431
tetragonal system, 648
tetraloop, 429, 645
tetramethylrhodamine (TMR), 516, 556
tetramethylsilane (TMS), 573
theorem
  binomial, 814, 877
  convolution, 903
  ergodic, 90, 453
  Nyquist-Shannon, 740, 752
  of minimal entropy production, 32
  projection slice, 758
  Pythagorean, 877
  Shannon sampling, 740
theory
  collision, 269
  Förster, 550
  transition state, 266
thermal energy
  and diffusion coefficient, 824
  at room temperature, 474
  definition, 268, 474
  in IR spectroscopy, 504
thermal equilibrium, 6, 29
thermal neutrons, 635
Thermal Shift Assay (TSA), 378
thermodynamic control, 270
thermodynamic cycle, 20, 76, 119, 353, 379
thermodynamic linkage, 78
thermodynamic stability, 40, 369, 490
thermodynamic temperature, 9
thermodynamics, 1
  of transport processes, 110
  statistical, 87, 95, 100
thermogenesis, 24
thermogenin, 118
thermogram, 864
thermoionic emission (of electrons), 653, 734
thermophoresis, 528
thiol coupling, 842
thioredoxin, 134, 317
third law of thermodynamics, 33
Thomson equation, 624
Thomson scattering, 615
Thomson scattering length, 625
Thon rings, 751
three-dimensional (3D) NMR, 592
threonine
  chiral side chain, 289
  glycosylation, 314
  phosphorylation, 315, 797
Tikhonov regularization, 608
tilt series, 756
time constant, 153
time of flight (TOF), 635, 804
time-correlated single photon counting (TCSPC), 543
time-resolved
  crystallography, 662, 679
  EPR, 603
  fluorescence, 542
  fluorescence anisotropy, 556
  FRET, 549, 560
  H/D exchange, 815
titration
  active site, 525
  and calorimetry, 853
  and EPR, 603
  and fluorescence, 523, 705
  and NMR, 574
  displacement, 526
  of polypeptides, 324
topoisomerase, 24, 418, 427
  and magnetic tweezers, 783
  and single-molecule FRET, 720
torsion angles, 290, 292
total internal reflection fluorescence (TIRF) microscopy, 697
trajectory
  analysis, 453
  in molecular dynamics, 449
  of diffusion, 113
  of ions in quadrupoles, 806
  single molecule tracking, 713
transducin, 483
transfer RNA (tRNA), 401, 431, 517
transformer, 454
transit time, 261
transition dipole moment, 476, 506
transition state
  analogs, 249
  and catalysis, 276, 281
  and isotope effects, 192
  in protein folding, 382
  stabilization by enzymes, 276
transition state theory, 266
transitions (allowed and forbidden), 476, 578
translation function, 671
translation vector, 650, 670

transmission coefficient, 268
transmission electron microscopy (TEM), 731, 733, 746
transport processes, 118
transverse relaxation optimized spectroscopy (TROSY), 594
transverse relaxation time, 576, 581, 586, 600, 603
trapping force, 776
triclinic system, 648
triglycerides, 354
trigonal system, 648
trigonometry, 877
trimolecular reaction, 147
triple helix (triplex)
  collagen, 316, 332
  in DNA, 412
  in riboswitches, 436
triplet state, 510
  magnetic moment for EPR, 599
  quenching, 710
trypsin, 281, 349, 815
tungsten lamp, 480, 508
turnover number, 218
turns
  in proteins, 337
  in RNA, 429
twisting number, 418
two-dimensional (2D) NMR, 585
two-photon excitation, 706, 723
two-state model (of protein folding), 374, 867
two-state system, 473
two-step binding, 194
  rate constant and amplitude, 200
tyrosine
  phosphorylation, 797
  tyrosyl radical, 601

U

ubiquinone, 116, 134
ultracentrifugation, 820
  principle, 821
  sedimentation equilibrium, 831
  sedimentation velocity, 826
  zonal, band, or isopycnic, 834
uncompetitive inhibition, 238
undulator, 656
unfolding
  measured by AFM, 778
  measured by DSC, 864
  measured by relaxation kinetics, 798
  of membrane proteins, 774
  of nucleic acids, 770, 779
  of proteins, 369
  of α-helices, 383
unimolecular reaction, 147
unit cell, 647
  and crystal lattice, 648
  and Miller indices, 658
  constants, 648
  motifs in, 649
  symmetry, 650
  vectors, 648
  volume of, 651
unusual nucleobases, 399, 400
UV damage, 483

V

van der Waals interactions, 303
  contact distance, 304, 408, 411
  geometries, 304
  in force fields, 444
  in ligand binding, 304
  in membrane formation, 304, 356
  in nucleic acids, 407, 411
  in protein stability, 369
  in protein-DNA interactions, 426
van der Waals radii, 303
van't Hoff enthalpy, 867
van't Hoff equation, 70
van't Hoff plot, 70
  linearity and heat capacity, 858
vanishing gradient, 454
vapor diffusion (for crystallization), 643
vapor pressure, 54
variance, 676, 895
vector
  between atoms in Patterson function, 664
  between transition dipoles, 552
  defining crystal lattice, 648
  defining reciprocal lattice, 659
  defining torsion angle, 292
  definition of, 886
  in circularly polarized light, 496
  in dipolar coupling, 597, 606
  in dipoles, 296
  in electromagnetic radiation, 471, 615
  in electron scattering, 753
  in energy minimization, 445
  in force measurements, 768
  in Harker construction, 666
  in light scattering, 628
  in linearly polarized light, 493
  nuclear spins, 580, 606
  representation of complex numbers, 893
  scalar, dot, cross, and triple product, 887
velocity Verlet algorithm, 452
Verlet integrator algorithm, 452
vertical transition, 476, *See* Franck-Frank-Condon principle
vesicle, 362
vibration frequency, 268
vibration spectroscopy, 504
vibrational transition, 623
vibronic relaxation (VR), 510
viscosity, 548
  and frictional coefficient, 824
  and sedimentation coefficient, 828
vision, 483
vitamin A, 355, *483*, *See also* retinal *and* carotenes
vitamin C (ascorbic acid), 130, 316, 388
vitamin E, 522
vitamin K, 130, 317
Viterbi path, 714
voltage-gated channel, 118

W

waiting time, 204
wall effect (in ultracentrifugation), 821
water
  dissociation of acids and bases in, 75
  electron density, 673
  exclusion from molecular interfaces, 347
  fluorescence quenching by, 529
  in DNA, 411, 427
  in hydrophobic environments, 370
  in macromolecular crystals, 641
  in membrane formation, 361
  in molecular dynamics simulations, 450
  in molecular interfaces, 347
  in protein folding, 369, 382

in RNA folding, 436
ionic product, 76
partial charges of, 299
rearrangement in absorption, 479
role in the hydrophobic effect, 349, 356
scattering densities, 627
solvent relaxation, 511
vibration, 506
Watson-Crick base pairs, 397
Watt, 854
wavelength, 472
and energy, 472, 506
de Broglie, 635
dependence of Rayleigh scattering on, 618
of electrons, 734
of EPR transitions, 600
of infrared light, 504
of neutrons, 635
of X-rays, 652, 655
wavenumbers (relation to wavelength), 506
wave-particle dualism, 635, 734
weak phase object, 750
weight-averaged sedimentation coefficient, 830
weights
in machine learning, 455
statistical, 88, 90, 96
wide-field microscopy, 693
wiggler, 656
Wiseman constant, 856, 860
wobble hypothesis, 401
work, 6, 20
electrical, 128
expansion, 6
work function, *See* also ionization energy
worm-like chain (WLC) model, 771

X

xanthophylls, 483
xenon arc lamp, 480, 512, 822
X-Gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside), 487
X-ray crystallography, 652
crystallization of macromolecules, 641
electron density and model building, 673
model refinement and validation, 675
phasing methods, 665
X-ray diffractometer, 661
X-rays, 472
detection, 662
diffraction, 657
generation, 653
scattering, 627, 632

Y

Y base (fluorescence), 515
yellow fluorescent protein (YFP), 520

Z

Zeeman effect, 570, 601
zero point, 22
zeroth law of thermodynamics, 6
Z-form (of DNA), 409
Zimm-Bragg model (of protein folding), 383
zinc (in electrochemistry), 127
zinc finger, 319, 424
zinc finger nucleases, 424
zinc hydrolase, 320
zonal centrifugation, 834

Prefixes.

name	prefix	scale	name	prefix	scale
quecto	q	10^{-30}	quetta	Q	10^{30}
ronto	r	10^{-27}	ronna	R	10^{27}
yocto	y	10^{-24}	yotta	Y	10^{24}
zepto	z	10^{-21}	zetta	Z	10^{21}
atto	a	10^{-18}	exa	E	10^{18}
femto	f	10^{-15}	peta	P	10^{15}
pico	p	10^{-12}	tera	T	10^{12}
nano	n	10^{-9}	giga	G	10^{9}
micro	μ	10^{-6}	mega	M	10^{6}
milli	m	10^{-3}	kilo	k	10^{3}
centi	c	10^{-2}	hecto	h	10^{2}
deci	d	10^{-1}	deka	da	10^{1}

The prefixes ranging from 10^{-30} to 10^{30} listed span approximately the orders of magnitudes one might encounter in biological systems. For instance, a single molecule in a liter of water formally has a concentration of 1.7 yM (yoctomolar), while the mass of a proton is about 1.7 yg (yoctogram). At the other end of the spectrum, we arrive at the yotta prefix when we consider the Avogadro constant of $6.022 \cdot 10^{23}$ particles per mol, which for example equals the number of water molecules in 18 mL. (The inverse of the Avogadro constant is also jokingly known as a "guacamole"). In 2022, two more prefixes were added, mainly to describe large amounts of data in informatics as ronna- and quetta-bytes. The tiny complements find application in particle physics and chemistry. For example, the mass of an electron is about 1 rg (rontogram).

SI (Système International) or base units.

Name	unit	definition
length	m	meter
mass	kg	kilogram
time	s	second
electric current	A	Ampère
temperature	K	Kelvin
amount	mol	mole
luminous intensity	cd	candela

There are only seven SI or base units. Note that the kilogram is the only SI unit that is defined with a prefix (kilo). Since 2019, all units are defined based on natural constants: the second is defined by a ^{133}Cs hyperfine transition frequency, the meter by the time light needs to travel, the kilogram *via* the Planck constant in $J \cdot s = kg \cdot m^2 \cdot s^{-1}$, the Ampère with the elementary charge expressed in A·s, the Kelvin by the Boltzmann constant expressed in $J \cdot K^{-1} = kg \cdot m^2 \cdot s^2 \cdot K^{-1}$, the mol by the Avogadro number, and the Candela by the luminous efficacy (not used in this book).

Derived units used in this book.

name	symbol	description	unit
°Celsius	°C	temperature	T (in K) −273.15 = T (in °C)
Coulomb	C	charge	A·s
Farad	F	electric capacitance	$A^2 \cdot s^4 \cdot kg^{-1} \cdot m^{-2}$
Gray	Gy	dose, energy per unit mass	$J \cdot kg^{-1}$
Hertz	Hz	frequency	s^{-1}
Joule	1 J = 1 Nm	energy, work, heat	$kg \cdot m^2 \cdot s^{-2}$
Joule per mole	1 J mol^{-1}	molar energy	$kg \cdot m^2 \cdot s^{-2} \cdot mol^{-1}$
Newton	N	force	$kg \cdot m \cdot s^{-2}$
Pascal	1 Pa = 1 N m^{-2}	pressure	$kg \cdot m^{-1} \cdot s^{-2}$
Radian	rad	angle	$m \cdot m^{-1}$
Tesla	T	magnetic field strength	$kg \cdot s^{-2} \cdot A^{-1}$
Volt	1 V = 1 W A^{-1}	voltage	$kg \cdot m^2 \cdot s^{-3} \cdot A^{-1}$
Watt	W = 1 J s^{-1}	power	$kg \cdot m^2 \cdot s^{-3}$

The voltage denominates an electrical potential difference. The Coulomb describes either the electric unit charge or a quantity of electricity. Some units are not based on the SI system but enjoy common use.

Non-SI derived units.

name	symbol	description	definition
Ångstrøm	Å	length	10^{-10} m
atmosphere	atm	pressure	101 325 Pa
bar	bar	pressure	10^{5} Pa
calorie	cal	energy	4.184 J
Dalton	Da	molecular mass	$1.6605 \cdot 10^{-27}$ kg
Debye	D	dipole moment	$3.33564 \cdot 10^{-30}$ C·m
electron volt	eV	energy	$1.602 \cdot 10^{-19}$ J
inch	In	length	0.0254 m
liter	L	volume	10^{-3} m^3
mile	mi	length	1.609.344 m
molar	M	concentration	mol·L^{-1}

While these units are not recommended, some of them are useful because they result in numbers that are easy to calculate with (compare for instance standard atmospheric pressure of 1 bar with the SI-based 10^5 Pa). The Da, L, and M are also used in this book. The Dalton (1 Da) is 1/12th of the mass of carbon ^{12}C.

Natural constants used in this book.

name	symbol	value
	universal	
gravitational acceleration	g	$9.80665\ \mathrm{m \cdot s^{-2}} \approx 9.81\ \mathrm{m\ s^{-2}}$
speed of light in vacuum	c	$299\,792\,458\ \mathrm{m \cdot s^{-1}} \approx 3 \cdot 10^{8}\ \mathrm{m\ s^{-1}}$
Planck constant	h	$6.62607015 \cdot 10^{-34}$ J·s
reduced Planck constant	$\hbar = h/(2\pi)$	$1.05457 \cdot 10^{-34}$ J·s
	electromagnetic	
vacuum permeability	μ_0	$4\pi \cdot 10^{-7}\ \mathrm{N \cdot A^{-2}} \approx 1.2566\ \mu\mathrm{N \cdot A^{-2}}$
vacuum permittivity	$\varepsilon_0 = 1/(\mu_0 c^2)_0$	$8.85419 \cdot 10^{-12}\ \mathrm{F \cdot m^{-1}} \approx 8.854\ \mathrm{pF \cdot m^{-1}}$
elementary charge	e	$1.602176634 \cdot 10^{-19}$ C
Bohr magneton	$\mu_B = e\hbar/(2m_e)$	$9.27401 \cdot 10^{-24}\ \mathrm{J \cdot T^{-1}}$
nuclear magneton	$\mu_N = e\hbar/(2m_p)$	$5.05078 \cdot 10^{-27}\ \mathrm{J \cdot T^{-1}}$
	atomic and nuclear	
electron mass	m_e	$9.10938 \cdot 10^{-31}$ kg
proton mass	m_p	$1.67262 \cdot 10^{-27}$ kg
neutron mass	m_n	$1.67493 \cdot 10^{-27}$ kg
	physico-chemical	
atomic mass	m_u	$1.66054 \cdot 10^{-27}$ kg
Avogadro's number	N_A	$6.02214076 \cdot 10^{23}\ \mathrm{mol^{-1}}$
Boltzmann constant	k_B	$1.380649 \cdot 10^{-23}\ \mathrm{J \cdot K^{-1}}$
Faraday constant	$F = N_A \cdot e$	$96\,485.3365\ \mathrm{C \cdot mol^{-1}}$
general gas constant	$R = k_B \cdot N_A$	$8.31446\ \mathrm{J \cdot mol^{-1} \cdot K^{-1}}$

The molar volume of an ideal gas at $T = 273.15$ K and $p = 101\,325$ Pa (1 atm pressure) is 22.413968 L mol^{-1}.