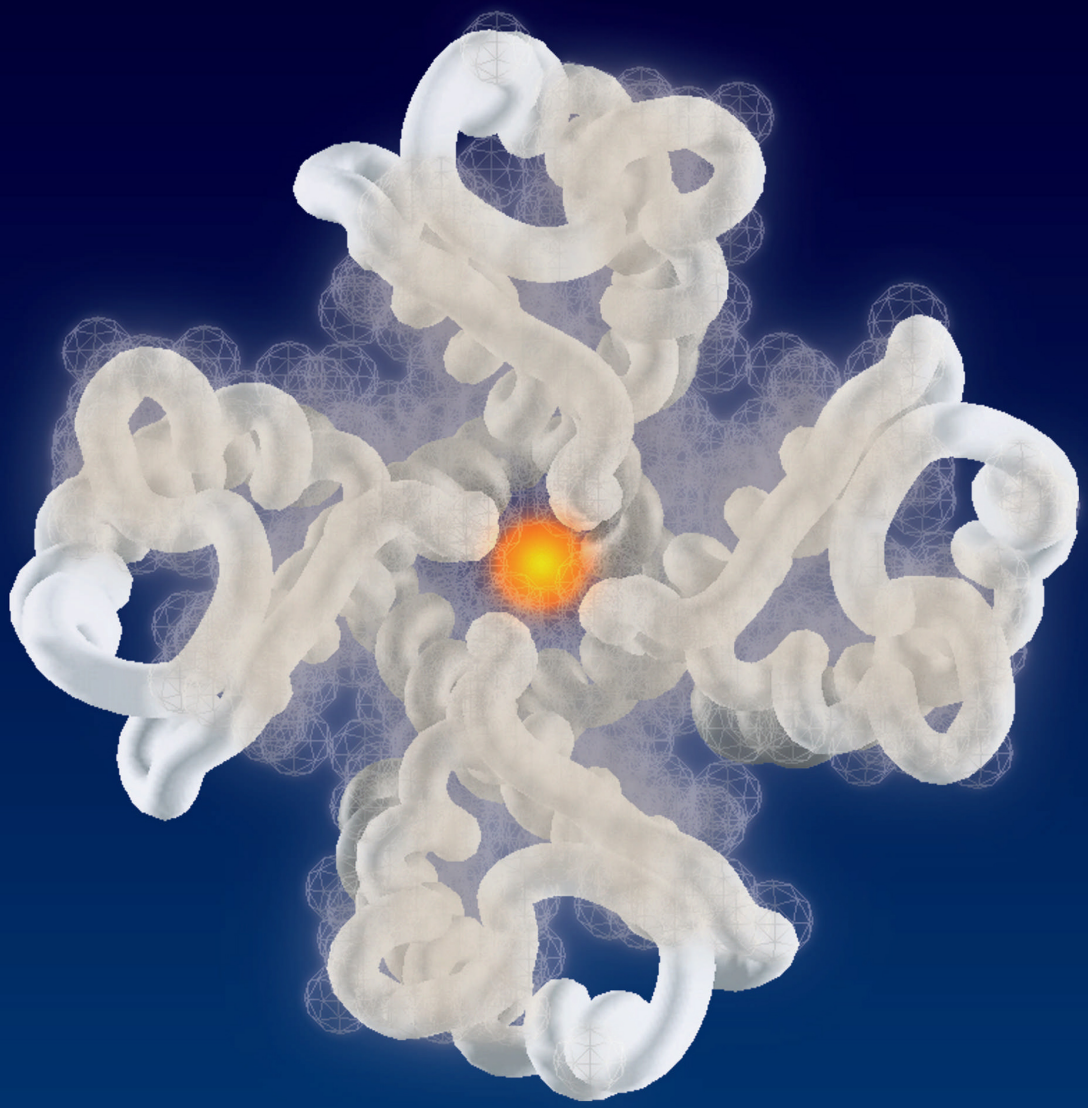
Introduction to Protein Structure

Second Edition



Carl Branden & John Tooze

Introduction to Protein Structure

Second Edition

An aerial view of the European Synchrotron Radiation Facility at Grenoble, France, an advanced source of synchrotron x-ray radiation for use in the study of protein structure, as well as for use in the physical and material sciences. The synchrotron radiation is produced in the circular building in the lower left of the photograph. (Courtesy of ESRF.)

Introduction to Protein Structure

Second Edition

Carl Branden

Microbiology and Tumor Biology Center
Karolinska Institute
Stockholm
Sweden

John Tooze

Imperial Cancer Research Fund Laboratories
Lincolns Inn Fields
London
UK

GARLAND PUBLISHING
ALERE FLAMMAM
Taylor & Francis Group

THE COVER

Front: The structure of the potassium channel from *Streptomyces lividans*, determined by Roderick MacKinnon at the Rockefeller University, New York. As discussed in Chapter 12, this structure—the first of such an ion channel—shows how the channel allows the passage of potassium ions through cell membranes with high efficiency and selectivity. The view is looking down the protein as it sits in the cell membrane, as seen from outside the cell, with a potassium ion shown in gold. This image was produced using the GRASP program (A. Nicholls and B. Honig, Columbia University) from atomic coordinates kindly provided by Rodney MacKinnon.
Back: A hand-drawn image of the potassium channel, in the same view as on the front cover, with each subunit of the tetrameric protein shown in a different color.
Cover design by Christopher Thorpe and Nigel Orme.

© 1991, 1999 Carl Branden and John Tooze

All rights reserved. No part of this book covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without permission of the publisher.

Visit the *Introduction to Protein Structure* Web site:
http://www.proteinstructure.com/

For information on other textbooks available from Garland Publishing,
visit: http://www.garlandpub.com/

Library of Congress Cataloging-in-Publication Data
Brändén, Carl-Ivar, 1934-
Introduction to protein structure / Carl Branden & John Tooze. -- 2nd ed.
p. ; cm.
Includes bibliographical references.
ISBN 978-0-8153-2305-1 (pbk.)
1. Proteins--Structure. I. Tooze, John. II. Title.
[DNLM: 1. Amino Acid Sequence. 2. Macromolecular Substances. 3. Molecular Structure. 4. Protein Conformation. 5. Proteins. QU 60 B819i 1999a]
QP551.B7635 2009
572'.633--dc22
2009017204

Published by Garland Science, Taylor & Francis Group, LLC,
an informa business,
711 Third Avenue, 8th floor, New York, NY 10017, USA,
and 3 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

15 14 13 12 11 10 9 8

Preface

The determination of the atomic structures of proteins has seen an enormous increase in impetus since the first edition of this book was published in 1991. The number of new structures reported is close to doubling each year. Technical advances—for example the increased availability of synchrotron x-ray beams and methods for freezing crystals so as to reduce radiation damage to them, the development of multidimensional NMR and NMR machines with ever more powerful magnetic fields, and the exploitation of gene cloning, sequencing and expression systems have all contributed to the growth of protein structure determination. On the one hand, it is becoming increasingly easy to obtain relatively large amounts of naturally rare proteins, on the other hand the crystallographers can work with ever smaller crystals.

The fundamental tenet of molecular biology, namely that one cannot really understand biological reactions without understanding the structure of the participating molecules, is at last being vindicated. As the database of known protein structures rapidly expands, so does the range of biological pathways about which we can ask meaningful questions at close to atomic levels of resolution. An understanding of the principles of protein structure is becoming of ever widening significance to molecular biology.

The pharmaceutical industry has over the past decade become a major user of the protein structure databases, and a major contributor of newly determined structures. Knowledge of an enzyme's or a receptor's atomic structure is invaluable in the search for specific and strongly binding inhibitors. For example the quest for effective inhibitors of HIV protease, to be used in combination therapy for AIDS, led many pharmaceutical companies to determine the structure of that protease with bound inhibitors. Over 120 of these structure determinations have been done so far and at least two inhibitors of HIV protease are now being regularly used to treat AIDS. It seems certain that the determination of the atomic structure of target molecules will play an increasingly important role in drug design.

The commercial exploitation of our increased understanding of protein structure will not, of course, be restricted to the pharmaceutical industry. The industrial use of enzymes in the chemical industry, the development of new and more specific pesticides and herbicides, the modification of enzymes in order to change the composition of plant oils and plant carbohydrates are all examples of other commercial developments that depend, in part, on understanding the structure of particular proteins at high resolution.

As the complete genomes of more and more species are sequenced, the determination of the function of previously unidentified open reading frames is becoming an increasing and challenging problem. The possibility of

Petitioner Merck, Ex. 1014, p. v

setting up centers for automated high through-put structure determinations is being seriously discussed. In the absence of any recognizable sequence homology to proteins of known function, this approach, surprising though it may seem, could become an effective way of determining function via structural homology.

The growth in the interest in high-resolution protein structure over the past decade and the reception of the first edition have encouraged us to prepare a new edition of this book. Universities are devoting more time to courses specifically on protein structure, or increasing the amount of time given to protein structure in more generally based biology and biochemistry courses. We hope that this new edition of *Introduction to Protein Structure* will prove useful both to teachers and students.

In 1988 when we began writing the first edition, about 250 protein structures had been determined to medium to high resolution and in those days a professional protein crystallographer was familiar with most of them. We were not therefore faced with a severe problem of what to leave out as we wrote. Today, the coordinates of over 6500 proteins have been deposited in the Protein Data Bank at Brookhaven, New York. Both the number of structures and the variety of biological systems to which they relate are so high that the field of protein structure is becoming more fragmented and specialized. It is becoming increasingly difficult to keep sight of the wood amongst so many trees. The question of what to include and what to omit is, for today's authors, crucially important. We have tried to resist the temptation to describe more and more proteins, adding detail but not increasing understanding of the basic concepts. This edition is inescapably a little larger than its predecessor, but to contain the increase in size we have deleted two chapters while adding four. We run the risk of disappointing not a few structural biologists whose favourite proteins are not mentioned. To them we apologize and ask for their understanding.

Petitioner Merck, Ex. 1014, p. vi

Acknowledgements

In preparing the second edition of this book we have again relied heavily on and benefited greatly from the advice and constructive criticism of numerous colleagues. We are particularly grateful to Ken Holmes (Max-Planck Institute, Heidelberg), Lawrence Stern (MIT), Michelle Arkin (Sunesis Pharmaceuticals), and Watson Fuller (Keele University, UK) for their contributions to, respectively, Chapters 14 and 18, Chapter 15, Chapter 17 and Chapter 18. Stephen Harrison (Harvard University) and Paul Sigler (Yale University) provided extensive help and advice on Chapters 8–10, 13 and 16, and Chapters 8–10 and 13, respectively, for which we are especially grateful.

The following, in alphabetical order, have reviewed one or more chapters, correcting our errors of fact or interpretation and helping to ensure they have the appropriate balance and emphasis: Tom Alber (University of California, Berkeley), Tom Blundell (Cambridge University, UK), Stephen Burley (Rockefeller University), Charles Craik (University of California, San Francisco), Ken Dill (University of California, San Francisco), Chris Dobson (Oxford University, UK), Anthony Fink (Unversity of California, Santa Cruz), Robert Fletterick (University of California, San Francisco), Richard Henderson (LMB, Cambridge, UK), Werner Kühlbrandt (MPI, Frankfurt), David Parry (Massey University, New Zealand), Greg Petsko (Brandeis University), and David Trentham (NIMR, London, UK).

The book depends for its accessibility upon its illustrations and we are hugely indebted to Nigel Orme, who, as with the first edition, has converted sketches into lucid figures. Keith Roberts has again advised us on how best graphically to represent chemical and structural phenomena. Jane Richardson (Duke University) has generously produced the Kinemage supplement to this edition and the book relies upon Richardson-type diagrams throughout to render the structures discussed comprehensible. We thank our publishers Garland Publishing, now part of the Taylor and Francis Group, for their support, and in particular Matthew Day for his enthusiastic editing of the complete manuscript. Miranda Robertson, in her inimitable style, has again managed the entire project.

Petitioner Merck, Ex. 1014, p. vii

Contents

Part I Basic Structural Principles 1

1. The Building Blocks 3
Proteins are polypeptide chains 4
The genetic code specifies 20 different amino acid side chains 4
Cysteines can form disulfide bridges 5
Peptide units are building blocks of protein structures 8
Glycine residues can adopt many different conformations 9
Certain side-chain conformations are energetically favorable 10
Many proteins contain intrinsic metal atoms 11
Conclusion 12
Selected readings 12

2. Motifs of Protein Structure 13
The interior of proteins is hydrophobic 14
The alpha (α) helix is an important element of secondary structure 14
The α helix has a dipole moment 16
Some amino acids are preferred in α helices 16
Beta (β) sheets usually have their β strands either parallel or antiparallel 19
Loop regions are at the surface of protein molecules 21
Schematic pictures of proteins highlight secondary structure 22
Topology diagrams are useful for classification of protein structures 23
Secondary structure elements are connected to form simple motifs 24
The hairpin β motif occurs frequently in protein structures 26
The Greek key motif is found in antiparallel β sheets 27
The β-α-β motif contains two parallel β strands 27
Protein molecules are organized in a structural hierarchy 28
Large polypeptide chains fold into several domains 29
Domains are built from structural motifs 30
Simple motifs combine to form complex motifs 30
Protein structures can be divided into three main classes 31
Conclusion 32
Selected readings 33

3. Alpha-Domain Structures 35
Coiled-coil α helices contain a repetitive heptad amino acid sequence pattern 35
The four-helix bundle is a common domain structure in α proteins 37
Alpha-helical domains are sometimes large and complex 39
The globin fold is present in myoglobin and hemoglobin 40
Geometric considerations determine α-helix packing 40
Ridges of one α helix fit into grooves of an adjacent helix 40
The globin fold has been preserved during evolution 41
The hydrophobic interior is preserved 42
Helix movements accommodate interior side-chain mutations 43
Sickle-cell hemoglobin confers resistance to malaria 43
Conclusion 45
Selected readings 45

4. Alpha/Beta Structures 47
Parallel β strands are arranged in barrels or sheets 47
Alpha/beta barrels occur in many different enzymes 48
Branched hydrophobic side chains dominate the core of α/β barrels 49
Pyruvate kinase contains several domains, one of which is an α/β barrel 51
Double barrels have occurred by gene fusion 52
The active site is formed by loops at one end of the α/β barrel 53

Petitioner Merck, Ex. 1014, p. ix

Alpha/beta barrels provide examples of evolution of new enzyme activities 54
Leucine-rich motifs form an α/β-horseshoe fold 55
Alpha/beta twisted open-sheet structures contain α helices on both sides of the β sheet 56
Open β-sheet structures have a variety of topologies 57
The positions of active sites can be predicted in α/β structures 57
Tyrosyl-tRNA synthetase has two different domains (α/β + α) 59
Carboxypeptidase is an α/β protein with a mixed β sheet 60
Arabinose-binding protein has two similar α/β domains 62
Conclusion 63
Selected readings 64

5. Beta Structures 67
Up-and-down barrels have a simple topology 68
The retinol-binding protein binds retinol inside an up-and-down β barrel 68
Amino acid sequence reflects β structure 69
The retinol-binding protein belongs to a superfamily of protein structures 70
Neuraminidase folds into up-and-down β sheets 70
Folding motifs form a propeller-like structure in neuraminidase 71
The active site is in the middle of one side of the propeller 72
Greek key motifs occur frequently in antiparallel β structures 72
The γ-crystallin molecule has two domains 74
The domain structure has a simple topology 74
Two Greek key motifs form the domain 74
The two domains have identical topology 75
The two domains have similar structures 76
The Greek key motifs in γ crystallin are evolutionarily related 76
The Greek key motifs can form jelly roll barrels 77
The jelly roll motif is wrapped around a barrel 77
The jelly roll barrel is usually divided into two sheets 78
The functional hemagglutinin subunit has two polypeptide chains 79
The subunit structure is divided into a stem and a tip 79
The receptor binding site is formed by the jelly roll domain 80
Hemagglutinin acts as a membrane fusogen 80
The structure of hemagglutinin is affected by pH changes 81
Parallel β-helix domains have a novel fold 84
Conclusion 85
Selected readings 87

6. Folding and Flexibility 89
Globular proteins are only marginally stable 90
Kinetic factors are important for folding 91
Molten globules are intermediates in folding 92
Burying hydrophobic side chains is a key event 93
Both single and multiple folding pathways have been observed 93
Enzymes assist formation of proper disulfide bonds during folding 96
Isomerization of proline residues can be a rate-limiting step in protein folding 98
Proteins can fold or unfold inside chaperonins 99
GroEL is a cylindrical structure with a central channel in which newly synthesized polypeptides bind 100
GroES closes off one end of the GroEL cylinder 102
The GroEL–GroES complex binds and releases newly synthesized polypeptides in an ATP-dependent cycle 102
The folded state has a flexible structure 104
Conformational changes in a protein kinase are important for cell cycle regulation 105
Peptide binding to calmodulin induces a large interdomain movement 109
Serpins inhibit serine proteinases with a spring-loaded safety catch mechanism 110
Effector molecules switch allosteric proteins between R and T states 113
X-ray structures explain the allosteric properties of phosphofructokinase 114
Conclusion 117
Selected readings 119

7. DNA Structures 121
The DNA double helix is different in A- and B-DNA 121
The DNA helix has major and minor grooves 122
Z-DNA forms a zigzag pattern 123
B-DNA is the preferred conformation *in vivo* 124
Specific base sequences can be recognized in B-DNA 124
Conclusion 125
Selected readings 126

Part 2 Structure, Function, and Engineering 127

8. DNA Recognition in Procaryotes by Helix-Turn-Helix Motifs 129
A molecular mechanism for gene control 129
Repressor and Cro proteins operate a procaryotic genetic switch region 130
The x-ray structure of the complete lambda Cro protein is known 131
The x-ray structure of the DNA-binding domain of the lambda repressor is known 132
Both lambda Cro and repressor proteins have a specific DNA-binding motif 133
Model building predicts Cro–DNA interactions 134
Genetic studies agree with the structural model 135
The x-ray structure of DNA complexes with 434 Cro and repressor revealed novel features of protein–DNA interactions 136
The structures of 434 Cro and the 434 repressor DNA-binding domain are very similar 137
The proteins impose precise distortions on the B-DNA in the complexes 138
Sequence-specific protein–DNA interactions recognize operator regions 138

Petitioner Merck, Ex. 1014, p. x

Protein–DNA backbone interactions determine DNA conformation 139
Conformational changes of DNA are important for differential binding of repressor and Cro to different operator sites 140
The essence of phage repressor and Cro 141
DNA binding is regulated by allosteric control 142
The *trp* repressor forms a helix-turn-helix motif 142
A conformational change operates a functional switch 142
Lac repressor binds to both the major and minor grooves inducing a sharp bend in the DNA 143
CAP-induced DNA bending could activate transcription 146
Conclusion 147
Selected readings 148

9. DNA Recognition by Eucaryotic Transcription Factors 151
Transcription is activated by protein–protein interactions 152
The TATA box-binding protein is ubiquitous 153
The three-dimensional structures of TBP–TATA box complexes are known 154
A β sheet in TBP forms the DNA-binding site 154
TBP binds in the minor groove and induces large structural changes in DNA 155
The interaction area between TBP and the TATA box is mainly hydrophobic 157
Functional implications of the distortion of DNA by TBP 158
TFIIA and TFIIB bind to both TBP and DNA 159
Homeodomain proteins are involved in the development of many eucaryotic organisms 159
Monomers of homeodomain proteins bind to DNA through a helix-turn-helix motif 160
In vivo specificity of homeodomain transcription factors depends on interactions with other proteins 162
POU regions bind to DNA by two tandemly oriented helix-turn-helix motifs 164
Much remains to be learnt about the function of homeodomains *in vivo* 166
Understanding tumorigenic mutations 166
The monomeric p53 polypeptide chain is divided in three domains 167
The oligomerization domain forms tetramers 167
The DNA-binding domain of p53 is an antiparallel β barrel 168
Two loop regions and one α helix of p53 bind to DNA 169
Tumorigenic mutations occur mainly in three regions involved in DNA binding 170
Conclusions 172
Selected readings 172

10. Specific Transcription Factors Belong to a Few Families 175
Several different groups of zinc-containing motifs have been observed 176
The classic zinc fingers bind to DNA in tandem along the major groove 177
The finger region of the classic zinc finger motif interacts with DNA 178
Two zinc-containing motifs in the glucocorticoid receptor form one DNA-binding domain 181
A dimer of the glucocorticoid receptor binds to DNA 183
An α helix in the first zinc motif provides the specific protein–DNA interactions 184
Three residues in the recognition helix provide the sequence-specific interactions with DNA 184
The retinoid X receptor forms heterodimers that recognize tandem repeats with variable spacings 185
Yeast transcription factor GAL4 contains a binuclear zinc cluster in its DNA-binding domain 187
The zinc cluster regions of GAL4 bind at the two ends of the enhancer element 188
The linker region also contributes to DNA binding 189
DNA-binding site specificity among the C_6-zinc cluster family of transcription factors is achieved by the linker regions 190
Families of zinc-containing transcription factors bind to DNA in several different ways 191
Leucine zippers provide dimerization interactions for some eucaryotic transcription factors 191
The GCN4 basic region leucine zipper binds DNA as a dimer of two uninterrupted α helices 193
GCN4 binds to DNA with both specific and nonspecific contacts 194
The HLH motif is involved in homodimer and heterodimer associations 196
The α-helical basic region of the b/HLH motif binds in the major groove of DNA 198
The b/HLH/zip family of transcription factors have both HLH and leucine zipper dimerization motifs 199
Max and MyoD recognize the DNA HLH consensus sequence by different specific protein–DNA interactions 201
Conclusion 201
Selected readings 203

11. An Example of Enzyme Catalysis: Serine Proteinases 205
Proteinases form four functional families 205
The catalytic properties of enzymes are reflected in K_m and k_{cat} values 206
Enzymes decrease the activation energy of chemical reactions 206
Serine proteinases cleave peptide bonds by forming tetrahedral transition states 208
Four important structural features are required for the catalytic action of serine proteinases 209
Convergent evolution has produced two different serine proteinases with similar catalytic mechanisms 210
The chymotrypsin structure has two antiparallel β-barrel domains 210

Petitioner Merck, Ex. 1014, p. xi

The active site is formed by two loop regions from each domain 211
Did the chymotrypsin molecule evolve by gene duplication? 212
Different side chains in the substrate specificity pocket confer preferential cleavage 212
Engineered mutations in the substrate specificity pocket change the rate of catalysis 213
The Asp 189-Lys mutation in trypsin causes unexpected changes in substrate specificity 215
The structure of the serine proteinase subtilisin is of the α/β type 215
The active sites of subtilisin and chymotrypsin are similar 216
A structural anomaly in subtilisin has functional consequences 217
Transition-state stabilization in subtilisin is dissected by protein engineering 217
Catalysis occurs without a catalytic triad 217
Substrate molecules provide catalytic groups in substrate-assisted catalysis 218
Conclusion 219
Selected readings 220

12. Membrane Proteins 223
Membrane proteins are difficult to crystallize 224
Novel crystallization methods are being developed 224
Two-dimensional crystals of membrane proteins can be studied by electron microscopy 225
Bacteriorhodopsin contains seven transmembrane α helices 226
Bacteriorhodopsin is a light-driven proton pump 227
Porins form transmembrane channels by β strands 228
Porin channels are made by up and down β barrels 229
Each porin molecule has three channels 230
Ion channels combine ion selectivity with high levels of ion conductance 232
The K^+ channel is a tetrameric molecule with one ion pore in the interface between the four subunits 232
The ion pore has a narrow ion selectivity filter 233
The bacterial photosynthetic reaction center is built up from four different polypeptide chains and many pigments 234
The L, M, and H subunits have transmembrane α helices 236
The photosynthetic pigments are bound to the L and M subunits 237
Reaction centers convert light energy into electrical energy by electron flow through the membrane 239
Antenna pigment proteins assemble into multimeric light-harvesting particles 240
Chlorophyll molecules form circular rings in the light-harvesting complex LH2 241
The reaction center is surrounded by a ring of 16 antenna proteins of the light-harvesting complex LH1 242
Transmembrane α helices can be predicted from amino acid sequences 244
Hydrophobicity scales measure the degree of hydrophobicity of different amino acid side chains 245
Hydropathy plots identify transmembrane helices 245
Reaction center hydropathy plots agree with crystal structural data 246
Membrane lipids have no specific interaction with protein transmembrane α helices 246
Conclusion 247
Selected readings 248

13. Signal Transduction 251
G proteins are molecular amplifiers 252
Ras proteins and the catalytic domain of G_α have similar three-dimensional structures 254
G_α is activated by conformational changes of three switch regions 257
GTPases hydrolyze GTP through nucleophilic attack by a water molecule 259
The G_β subunit has a seven-blade propeller fold, built up from seven WD repeat units 261
The GTPase domain of G_α binds to G_β in the heterotrimeric $G_{\alpha\beta\gamma}$ complex 263
Phosducin regulates light adaptation in retinal rods 265
Phosducin binding to $G_{\beta\gamma}$ blocks binding of G_α 265
The human growth hormone induces dimerization of its cognate receptor 267
Dimerization of the growth hormone receptor is a sequential process 268
The growth hormone also binds to the prolactin receptor 269
Tyrosine kinase receptors are important enzyme-linked receptors 270
Small protein modules form adaptors for a signaling network 272
SH2 domains bind to phosphotyrosine-containing regions of target molecules 273
SH3 domains bind to proline-rich regions of target molecules 274
Src tyrosine kinases comprise SH2 and SH3 domains in addition to a tyrosine kinase 275
The two domains of the kinase in the inactive state are held in a closed conformation by assembly of the regulatory domains 277
Conclusion 278
Selected readings 280

14. Fibrous Proteins 283
Collagen is a superhelix formed by three parallel, very extended left-handed helices 284
Coiled coils are frequently used to form oligomers of fibrous and globular proteins 286
Amyloid fibrils are suggested to be built up from continuous β sheet helices 288
Spider silk is nature's high-performance fiber 289
Muscle fibers contain myosin and actin which slide against each other during muscle contraction 290

Petitioner Merck, Ex. 1014, p. xii

Myosin heads form cross-bridges between the actin and myosin filaments 291
Time-resolved x-ray diffraction of frog muscle confirmed movement of the cross-bridges 292
Structures of actin and myosin have been determined 293
The structure of myosin supports the swinging cross-bridge hypothesis 295
The role of ATP in muscular contraction has parallels to the role of GTP in G-protein activation 296
Conclusion 297
Selected readings 298

15. Recognition of Foreign Molecules by the Immune System 299
The polypeptide chains of antibodies are divided into domains 300
Antibody diversity is generated by several different mechanisms 302
All immunoglobulin domains have similar three-dimensional structures 303
The immunoglobulin fold is best described as two antiparallel β sheets packed tightly against each other 304
The hypervariable regions are clustered in loop regions at one end of the variable domain 305
The antigen-binding site is formed by close association of the hypervariable regions from both heavy and light chains 306
The antigen-binding site binds haptens in crevices and protein antigens on large flat surfaces 308
The CDR loops assume only a limited range of conformations, except for the heavy chain CDR3 311
An IgG molecule has several degrees of conformational flexibility 312
Structures of MHC molecules have provided insights into the molecular mechanisms of T-cell activation 312
MHC molecules are composed of antigen-binding and immunoglobulin-like domains 313
Recognition of antigen is different in MHC molecules compared with immunoglobulins 314
Peptides are bound differently by class I and class II MHC molecules 315
T-cell receptors have variable and constant immunoglobulin domains and hypervariable regions 316
MHC–peptide complexes are the ligands for T-cell receptors 318
Many cell-surface receptors contain immunoglobulin-like domains. 318
Conclusion 320
Selected readings 321

16. The Structure of Spherical Viruses 325
The protein shells of spherical viruses have icosahedral symmetry 327
The icosahedron has high symmetry 327
The simplest virus has a shell of 60 protein subunits 328
Complex spherical viruses have more than one polypeptide chain in the asymmetric unit 329
Structural versatility gives quasi-equivalent packing in T = 3 plant viruses 331
The protein subunits recognize specific parts of the RNA inside the shell 332
The protein capsid of picornaviruses contains four polypeptide chains 333
There are four different structural proteins in picornaviruses 334
The arrangement of subunits in the shell of picornaviruses is similar to that of T = 3 plant viruses 334
The coat proteins of many different spherical plant and animal viruses have similar jelly roll barrel structures, indicating an evolutionary relationship 335
Drugs against the common cold may be designed from the structure of rhinovirus 337
Bacteriophage MS2 has a different subunit structure 339
A dimer of MS2 subunits recognizes an RNA packaging signal 339
The core protein of alphavirus has a chymotrypsin-like fold 340
SV40 and polyomavirus shells are constructed from pentamers of the major coat protein with nonequivalent packing but largely equivalent interactions 341
Conclusion 343
Selected readings 344

17. Prediction, Engineering, and Design of Protein Structures 347
Homologous proteins have similar structure and function 348
Homologous proteins have conserved structural cores and variable loop regions 349
Knowledge of secondary structure is necessary for prediction of tertiary structure 350
Prediction methods for secondary structure benefit from multiple alignment of homologous proteins 351
Many different amino acid sequences give similar three-dimensional structures 352
Prediction of protein structure from sequence is an unsolved problem 352
Threading methods can assign amino acid sequences to known three-dimensional folds 353
Proteins can be made more stable by engineering 354
Disulfide bridges increase protein stability 355
Glycine and proline have opposite effects on stability 356
Stabilizing the dipoles of α helices increases stability 357
Mutants that fill cavities in hydrophobic cores do not stabilize T4 lysozyme 358
Proteins can be engineered by combinatorial methods 358
Phage display links the protein library to DNA 359
Affinity and specificity of proteinase inhibitors can be optimized by phage display 361

Petitioner Merck, Ex. 1014, p. xiii

Structural scaffolds can be reduced in size while function is retained 363
Phage display of random peptide libraries identified agonists of erythropoetin receptor 364
DNA shuffling allows accelerated evolution of genes 365
Protein structures can be designed from first principles 367
A β structure has been converted to an α structure by changing only half of the sequence 368
Conclusion 370
Selected readings 371

18. Determination of Protein Structures 373
Several different techniques are used to study the structure of protein molecules 373
Protein crystals are difficult to grow 374
X-ray sources are either monochromatic or polychromatic 376
X-ray data are recorded either on image plates or by electronic detectors 377
The rules for diffraction are given by Bragg's law 378
Phase determination is the major crystallographic problem 379
Phase information can also be obtained by Multiwavelength Anomalous Diffraction experiments 381
Building a model involves subjective interpretation of the data 381
Errors in the initial model are removed by refinement 383
Recent technological advances have greatly influenced protein crystallography 383
X-ray diffraction can be used to study the structure of fibers as well as crystals 384
The structure of biopolymers can be studied using fiber diffraction 386
NMR methods use the magnetic properties of atomic nuclei 387
Two-dimensional NMR spectra of proteins are interpreted by the method of sequential assignment 389
Distance constraints are used to derive possible structures of a protein molecule 390
Biochemical studies and molecular structure give complementary functional information 391
Conclusion 391
Selected readings 392

Protein Structure on the World Wide Web 393

Petitioner Merck, Ex. 1014, p. xiv

Part 1

Basic Structural Principles

Petitioner Merck, Ex. 1014, p. 1

The Building Blocks

1

Recombinant DNA techniques have provided tools for the rapid determination of DNA sequences and, by inference, the amino acid sequences of proteins from structural genes. The number of such sequences is now increasing almost exponentially, but by themselves these sequences tell little more about the biology of the system than a New York City telephone directory tells about the function and marvels of that city.

The proteins we observe in nature have evolved, through selective pressure, to perform specific functions. The functional properties of proteins depend upon their three-dimensional structures. The three-dimensional structure arises because particular sequences of amino acids in polypeptide chains fold to generate, from linear chains, compact domains with specific three-dimensional structures (Figure 1.1). The folded domains can serve as modules for building up large assemblies such as virus particles or muscle fibers, or they can provide specific catalytic or binding sites, as found in enzymes or proteins that carry oxygen or that regulate the function of DNA.

To understand the biological function of proteins we would therefore like to be able to deduce or predict the three-dimensional structure from the amino acid sequence. This we cannot do. In spite of considerable efforts over the past 25 years, this folding problem is still unsolved and remains one of the most basic intellectual challenges in molecular biology.

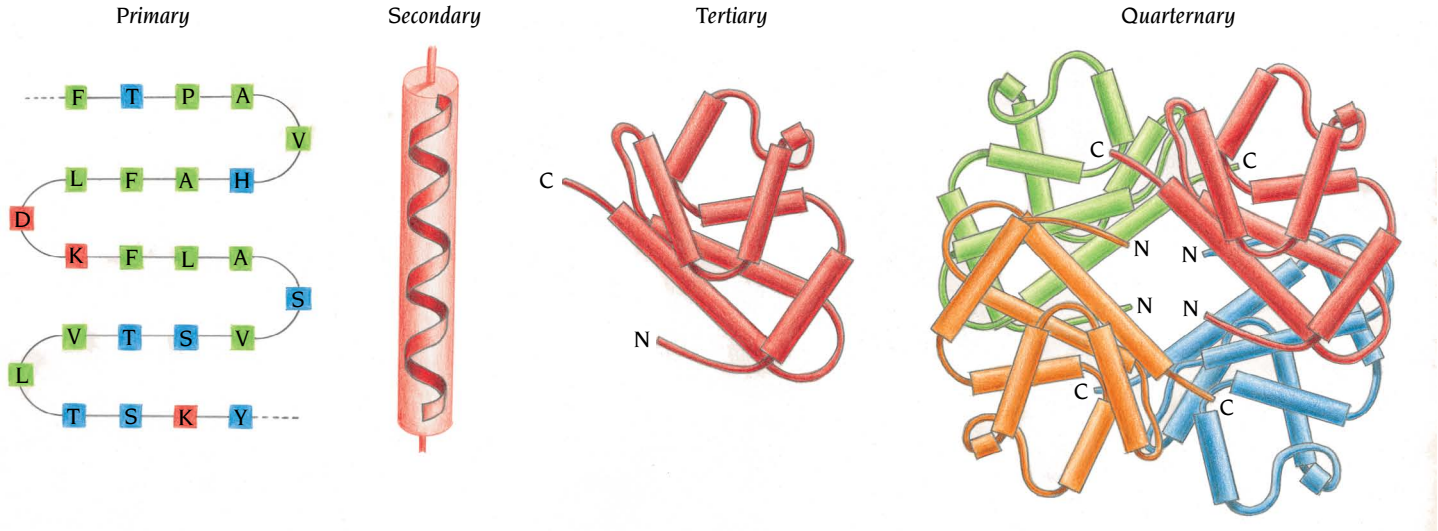


Figure 1.1 The amino acid sequence of a protein's polypeptide chain is called its **primary** structure. Different regions of the sequence form local regular **secondary** structures, such as alpha (α) helices or beta (β) strands. The **tertiary** structure is formed by packing such structural elements into one or several compact globular units called domains. The final protein may contain several polypeptide chains arranged in a **quaternary** structure. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, an **active site**.

Petitioner Merck, Ex. 1014, p. 3

Protein folding remains a problem because there are 20 different amino acids that can be combined into many more different proteins than there are atoms in the known universe. In addition there is a vast number of ways in which similar structural domains can be generated in proteins by different amino acid sequences. By contrast, the structure of DNA, made up of only four different nucleotide building blocks that occur in two pairs, is relatively simple, regular, and predictable.

Since the three-dimensional structures of individual proteins cannot be predicted, they must instead be determined experimentally by x-ray crystallography, electron crystallography or nuclear magnetic resonance (NMR) techniques. Over the past 30 years the structures of more than 6000 proteins have been solved, and the sequences of more than 500,000 have been determined. This has generated a body of information from which a set of basic principles of protein structure has emerged. These principles make it easier for us to understand how protein structure is generated, to identify common structural themes, to relate structure to function, and to see fundamental relationships between different proteins. The science of protein structure is at the stage of taxonomy where we can begin to discern patterns and motifs among the relatively small number of proteins whose three-dimensional structure is known.

The first six chapters of this book deal with the basic principles of protein structure as we understand them today, and examples of the different major classes of protein structures are presented. Chapter 7 contains a brief discussion on DNA structures with emphasis on recognition by proteins of specific nucleotide sequences. The remaining chapters illustrate how during evolution different structural solutions have been selected to fulfill particular functions.

Proteins are polypeptide chains

All of the 20 amino acids have in common a central carbon atom (C_α) to which are attached a hydrogen atom, an amino group (NH_2), and a carboxyl group (COOH) (Figure 1.2a). What distinguishes one amino acid from another is the side chain attached to the C_α through its fourth valence. There are 20 different side chains specified by the genetic code; others occur, in rare cases, as the products of enzymatic modifications after translation.

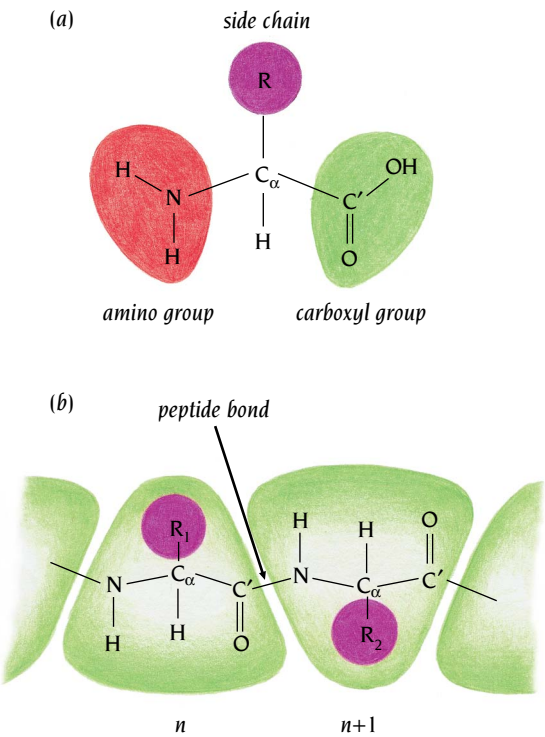
Amino acids are joined end-to-end during protein synthesis by the formation of **peptide bonds** when the carboxyl group of one amino acid condenses with the amino group of the next to eliminate water (Figure 1.2b). This process is repeated as the chain elongates. One consequence is that the amino group of the first amino acid of a polypeptide chain and the carboxyl group of the last amino acid remain intact, and the chain is said to extend from its amino terminus to its carboxy terminus. The formation of a succession of peptide bonds generates a "main chain," or "backbone," from which project the various side chains.

The main-chain atoms are a carbon atom C_α to which the side chain is attached, an NH group bound to C_α, and a carbonyl group C′=O, where the carbon atom C′ is attached to C_α. These units, or residues, are linked into a polypeptide by a peptide bond between the C′ atom of one residue and the nitrogen atom of the next (see Figure 1.2b). The basic repeating unit along the main chain from a biochemical or genetic viewpoint is thus (NH–C_αH–C′=O), which is the residue of the common parts of amino acids after peptide bonds have been formed (see Figure 1.2b).

The genetic code specifies 20 different amino acid side chains

The 20 different side chains that occur in proteins are shown in Panel 1.1 (pp. 6–7). Their names are abbreviated with both a three-letter and a one-letter code, which are also given in the panel. The one-letter codes are worth memorizing, as they are widely used in the literature. A mnemonic device for linking the one-letter code to the names of the amino acids is given in Panel 1.1.

Figure 1.2 Proteins are built up by amino acids that are linked by peptide bonds to form a polypeptide chain. (a) Schematic diagram of an amino acid, illustrating the nomenclature used in this book. A central carbon atom (C_α) is attached to an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom (H), and a side chain (R). (b) In a polypeptide chain the carboxyl group of amino acid *n* has formed a peptide bond, C–N, to the amino group of amino acid *n* + 1. One water molecule is eliminated in this process. The repeating units, which are called residues, are divided into main-chain atoms and side chains. The main-chain part, which is identical in all residues, contains a central C_α atom attached to an NH group, a C′=O group, and an H atom. The side chain R, which is different for different residues, is bound to the C_α atom.



Petitioner Merck, Ex. 1014, p. 4

The amino acids are usually divided into three different classes defined by the chemical nature of the side chain. The first class comprises those with strictly hydrophobic side chains: Ala (A), Val (V), Leu (L), Ile (I), Phe (F), Pro (P), and Met (M). The four charged residues, Asp (D), Glu (E), Lys (K), and Arg (R), form the second class. The third class comprises those with polar side chains: Ser (S), Thr (T), Cys (C), Asn (N), Gln (Q), His (H), Tyr (Y), and Trp (W). The amino acid glycine (G), which has only a hydrogen atom as a side chain and so is the simplest of the 20 amino acids, has special properties and is usually considered either to form a fourth class or to belong to the first class.

The four groups attached to the C_α atom are chemically different for all the amino acids except glycine, where two H atoms bind to C_α. All amino acids except glycine are thus chiral molecules that can exist in two different forms with different "hands," L- or D-form (Figure 1.3).

Biological systems depend on specific detailed recognition of molecules that distinguish between chiral forms. The translation machinery for protein synthesis has evolved to utilize only one of the chiral forms of amino acids, the L-form. All amino acids that occur in proteins therefore have the L-form. There is, however, no obvious reason why the L-form was chosen during evolution and not the D-form.

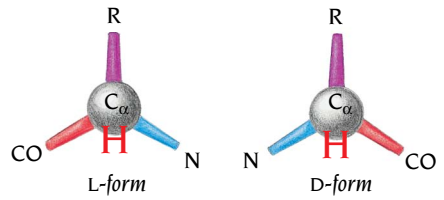


Figure 1.3 The "handedness" of amino acids. Looking down the H–C_α bond from the hydrogen atom, the L-form has CO, R, and N substituents from C_α going in a clockwise direction. There is a mnemonic to remember this; for the L-form the groups read CORN in clockwise direction.

Cysteines can form disulfide bridges

Two cysteine residues in different parts of the polypeptide chain but adjacent in the three-dimensional structure of a protein can be oxidized to form a **disulfide bridge** (Figure 1.4). The disulfide is usually the end product of air oxidation according to the following reaction scheme:

$$2\ \text{–}CH_2SH + \tfrac{1}{2}\ O_2 \rightleftharpoons \text{–}CH_2\text{–}S\text{–}S\text{–}CH_2 + H_2O$$

This reaction requires an oxidative environment, and such disulfide bridges are usually not found in intracellular proteins, which spend their lifetime in an essentially reductive environment. Disulfide bridges do, however, occur quite frequently among extracellular proteins that are secreted from cells, and in eucaryotes, formation of these bridges occurs within the lumen of the endoplasmic reticulum, the first compartment of the secretory pathway.

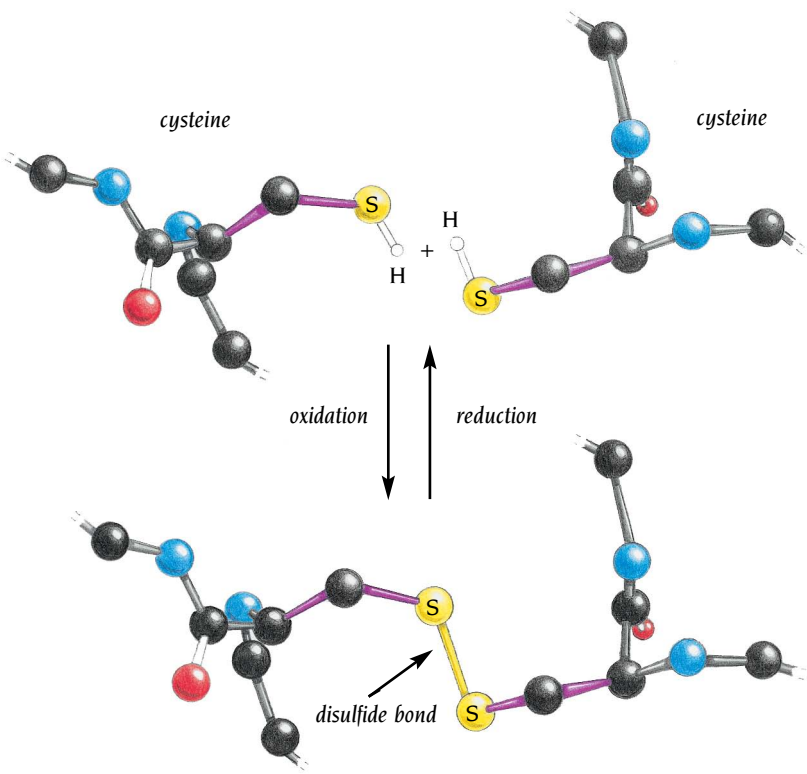


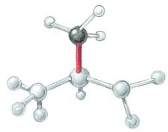
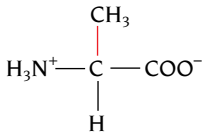
Figure 1.4 The disulfide is usually the end product of air oxidation according to the following schematic reaction scheme:

$$2\ \text{–}CH_2SH + \tfrac{1}{2}\ O_2 \rightleftharpoons \text{–}CH_2\text{–}S\text{–}S\text{–}CH_2 + H_2O$$

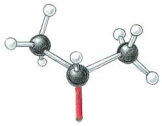
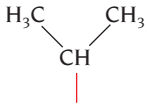
Disulfide bonds form between the side chains of two cysteine residues. Two SH groups from cysteine residues, which may be in different parts of the amino acid sequence but adjacent in the three-dimensional structure, are oxidized to form one S–S (disulfide) group.

Petitioner Merck, Ex. 1014, p. 5

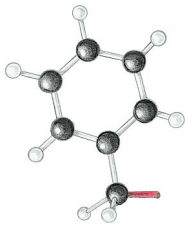
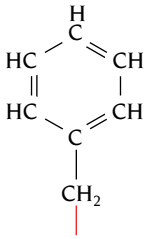
(a) Hydrophobic amino acids

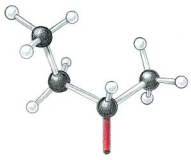
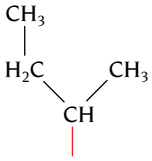
A Ala, Alanine

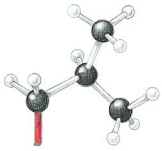
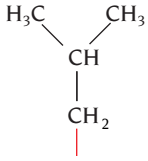
V Val, Valine

F Phe, Phenylalanine

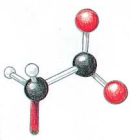
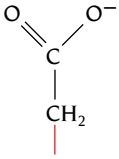



I Ile, Isoleucine

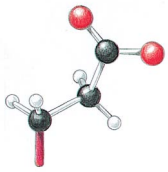
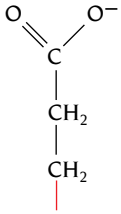



L Leu, Leucine

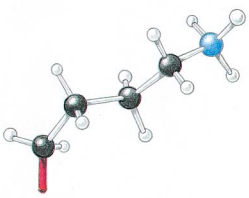
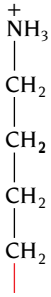
(b) Charged amino acids

D Asp, Aspartic acid

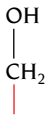



E Glu, Glutamic acid

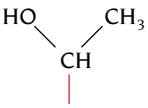



K Lys, Lysine

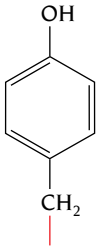
(c) Polar amino acids



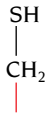
S Ser, Serine



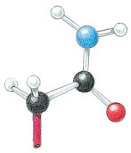
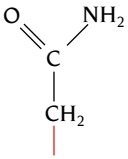
T Thr, Threonine



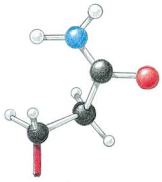
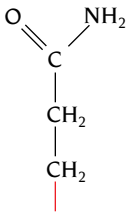
Y Tyr, Tyrosine



C Cys, Cysteine

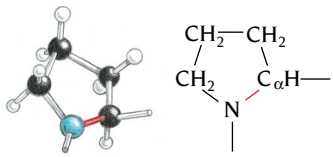



N Asn, Asparagine

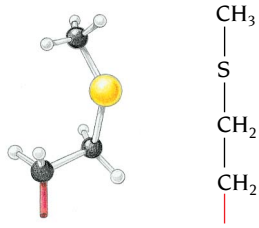



Q Gln, Glutamine

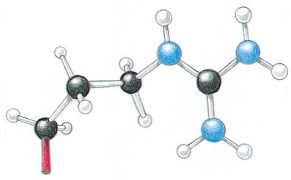
Petitioner Merck, Ex. 1014, p. 6



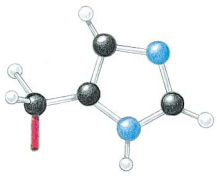
P *Pro, Proline*



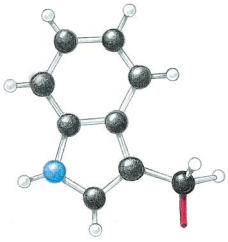
M *Met, Methionine*



R *Arg, Arginine*



H *His, Histidine*



W *Trp, Tryptophan*

Panel 1.1 The 20 different amino acids that occur in proteins. Only side chains are shown, except for the first amino acid, alanine, where all atoms are shown. The bond from the side chain to C_α is red. A ball-and-stick model, the chemical formula, the full name, and the three-letter and one-letter codes are given for each amino acid.

There are some easy ways of remembering the one-letter code for amino acids. If only one amino acid begins with a certain letter, that letter is used:

C = Cys = Cysteine
H = His = Histidine
I = Ile = Isoleucine
M = Met = Methionine
S = Ser = Serine
V = Val = Valine

If more than one amino acid begins with a certain letter, that letter is assigned to the most commonly occurring amino acid:

A = Ala = Alanine
G = Gly = Glycine
L = Leu = Leucine
P = Pro = Proline
T = Thr = Threonine

Some of the others are phonetically suggestive:

F = Phe = Phenylalanine (“Fenylalanine”)
R = Arg = Arginine (“aRginine”)
Y = Tyr = Tyrosine (“tYrosine”)
W = Trp = Tryptophan (double ring in the molecule)

In other cases a letter close to the initial is used. Amides have letters from the middle of the alphabet. The smaller molecules (D, N, B) are earlier in the alphabet than the larger ones (E, Q, Z).

D = Asp = Aspartic acid (near A)
N = Asn = Asparagine (contains N)
B = Asx = Either of D or N
E = Glu = Glutamic acid (near G)
Q = Gln = Glutamine (“Q-tamine”)
Z = Glx = Either of E or Q
K = Lys = Lysine (near L)
X = X = Undetermined amino acid

(*d*) *Glycine*

G *Gly, Glycine*

Petitioner Merck, Ex. 1014, p. 7

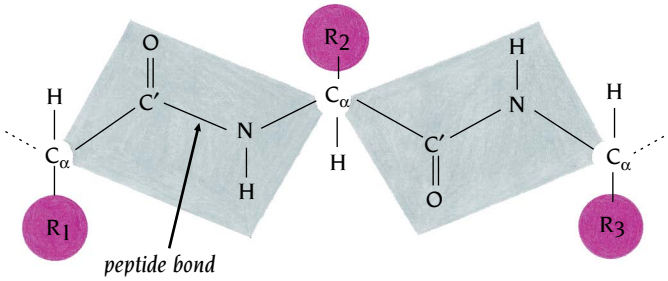


Figure 1.5 Part of a polypeptide chain that is divided into peptide units, represented as blocks in the diagram. Each peptide unit contains the C_α atom and the C′=O group of residue *n* as well as the NH group and the C_α atom of residue *n* + 1. Each such unit is a planar, rigid group with known bond distances and bond angles. R_1, R_2, and R_3 are the side chains attached to the C_α atoms that link the peptide units in the polypeptide chain. The peptide group is planar because the additional electron pair of the C=O bond is delocalized over the peptide group such that rotation around the C–N bond is prevented by an energy barrier.

Disulfide bridges stabilize three-dimensional structure. In some proteins these bridges hold together different polypeptide chains; for example, the A and B chains of insulin are linked by two disulfide bridges between the chains. More frequently intramolecular disulfide bridges stabilize the folding of a single polypeptide chain, making the protein less susceptible to degradation. There are many examples of this, including proteins with short polypeptide chains, such as snake venom toxins and protease inhibitors, that need additional stabilizing factors to produce a stable fold. Much effort is currently spent on introducing extra intramolecular disulfide bridges into enzymes by site-directed mutagenesis in order to make them more thermostable and hence more useful for industrial applications as catalysts, as described in Chapter 17.

Peptide units are building blocks of protein structures

Figure 1.2 shows one way of dividing a polypeptide chain, the biochemist's way. There is, however, a different way to divide the main chain into repeating units that is preferable when we want to describe the structural properties of proteins. For this purpose it is more useful to divide the polypeptide chain into peptide units that go from one C_α atom to the next C_α atom (see Figure 1.5). Each C_α atom, except the first and the last, thus belongs to two such units. The reason for dividing the chain in this way is that all the atoms in such a unit are fixed in a plane with the bond lengths and bond angles very nearly the same in all units in all proteins. Note that the peptide units of the main chain do not involve the different side chains (Figure 1.5). We will use both of these alternative descriptions of polypeptide chains—the biochemical and the structural—and discuss proteins in terms of the sequence of different amino acids and the sequence of planar peptide units.

Since the peptide units are effectively rigid groups that are linked into a chain by covalent bonds at the C_α atoms, the only degrees of freedom they have are rotations around these bonds. Each unit can rotate around two such bonds: the C_α–C′ and the N–C_α bonds (Figure 1.6). By convention the angle of rotation around the N–C_α bond is called **phi (ϕ)** and the angle around the C_α–C′ bond from the same C_α atom is called **psi (ψ).**

In this way each amino acid residue is associated with two conformational angles ϕ and ψ. Since these are the only degrees of freedom, the conformation of the whole main chain of the polypeptide is completely determined when the ϕ and ψ angles for each amino acid are defined with high accuracy.

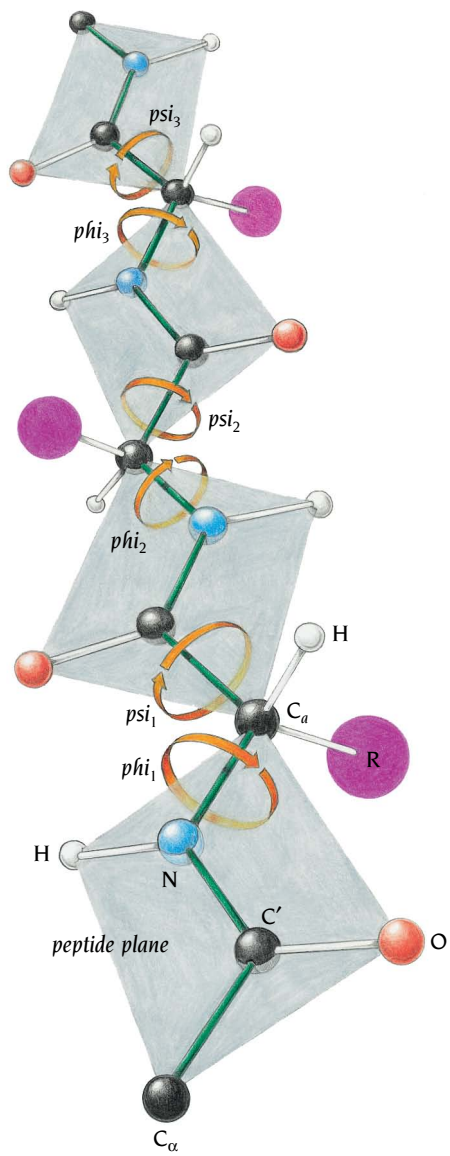


Figure 1.6 Diagram showing a polypeptide chain where the main-chain atoms are represented as rigid peptide units, linked through the C_α atoms. Each unit has two degrees of freedom; it can rotate around two bonds, its C_α–C′ bond and its N–C_α bond. The angle of rotation around the N–C_α bond is called phi (ϕ) and that around the C_α–C′ bond is called psi (ψ). The conformation of the main-chain atoms is therefore determined by the values of these two angles for each amino acid.

Petitioner Merck, Ex. 1014, p. 8

Glycine residues can adopt many different conformations

Most combinations of ϕ and ψ angles for an amino acid are not allowed because of steric collisions between the side chains and main chain. It is reasonably straightforward to calculate those combinations that are allowed. Since the D- and L-forms of the amino acids have their side chain oriented differently with respect to the CO group, they have different allowed ϕ and ψ angles. Proteins built from D-amino acids would thus be expected to have different conformations from those found in nature that are exclusively made of L-amino acids. Since the L- and D-forms of each amino acid are mirror images of one another, would a protein made exclusively of D-form residues produce a structure that is the mirror image of the natural protein? Stephen Kent and his colleagues at the Scripps Institute chemically synthesized both the L- and the D-forms of HIV-1 protease. The D-enzyme proved indeed to be the mirror image of the L-enzyme. Furthermore the D-enzyme and L-enzyme had reciprocal chiral specificity on peptide substrates, the D-enzyme only recognizing and cutting peptides made of D-amino acids. Perhaps the choice of the L-form at the outset of the evolution of life on earth was random and irrevocable.

The angle pairs ϕ and ψ are usually plotted against each other in a diagram called a **Ramachandran plot** after the Indian biophysicist G.N. Ramachandran who first made calculations of sterically allowed regions. Figure 1.7 shows the results of such calculations and also a plot for all amino

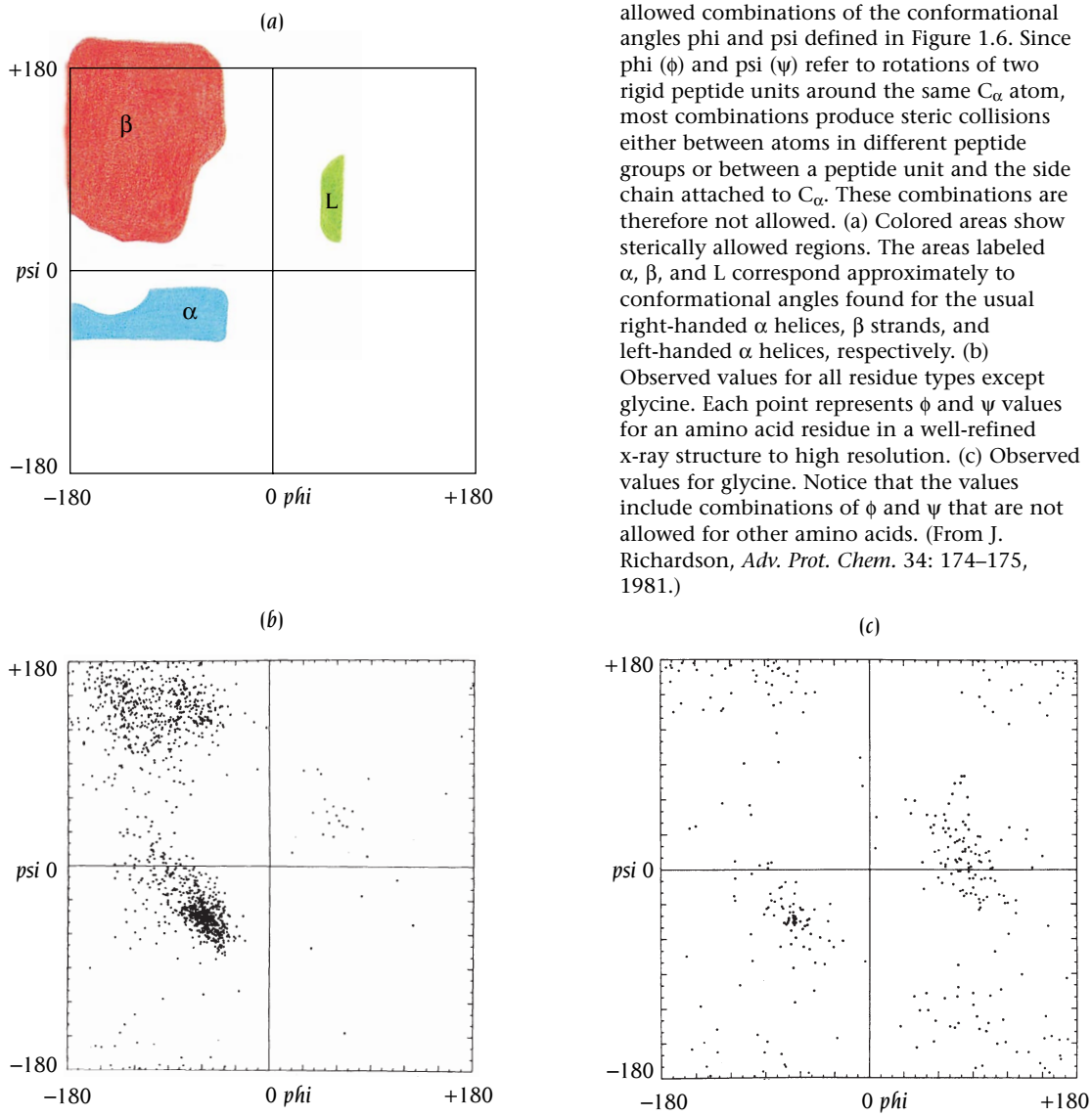


Figure 1.7 Ramachandran plots showing allowed combinations of the conformational angles phi and psi defined in Figure 1.6. Since phi (ϕ) and psi (ψ) refer to rotations of two rigid peptide units around the same C_α atom, most combinations produce steric collisions either between atoms in different peptide groups or between a peptide unit and the side chain attached to C_α. These combinations are therefore not allowed. (a) Colored areas show sterically allowed regions. The areas labeled α, β, and L correspond approximately to conformational angles found for the usual right-handed α helices, β strands, and left-handed α helices, respectively. (b) Observed values for all residue types except glycine. Each point represents ϕ and ψ values for an amino acid residue in a well-refined x-ray structure to high resolution. (c) Observed values for glycine. Notice that the values include combinations of ϕ and ψ that are not allowed for other amino acids. (From J. Richardson, *Adv. Prot. Chem.* 34: 174–175, 1981.)

Petitioner Merck, Ex. 1014, p. 9

acids except glycine from a number of accurately determined protein structures. It is apparent that the observed values are clustered in the sterically allowed regions. There is one important exception. Glycine, with only a hydrogen atom as a side chain, can adopt a much wider range of conformations than the other residues, as seen in Figure 1.7c. Glycine thus plays a structurally very important role; it allows unusual main-chain conformations in proteins. This is one of the main reasons why a high proportion of glycine residues are conserved among homologous protein sequences.

Regions in the Ramachandran plot are named after the conformation that results in a peptide if the corresponding ϕ and ψ angles are repeated in successive amino acids along the chain. The major allowed regions in Figure 1.7a are the right-handed α-helical cluster in the lower left quadrant (see Chapter 3); the broad region of extended β strands of both parallel and antiparallel β structures (see Chapter 4) in the upper left quadrant; and the small, sparsely populated left-handed α-helical region in the upper right quadrant. Left-handed α helices are usually found in loop regions or in small single-turn α helices.

Certain side-chain conformations are energetically favorable

Any side chain longer than that of alanine can in principle have several different conformations because of rotations around the bonds between the side-chain carbon atoms. It is a general rule in chemistry that the most energetically favored arrangements for two tetrahedrally coordinated carbon atoms are the "staggered" conformations, in which the substituents of one carbon atom are between those of the other when viewed along the axis of rotation, as illustrated in Figure 1.8. For each such carbon atom in a side chain, there are three possible staggered conformations, which are related by a three step, 120° rotation around the carbon–carbon bond. These three conformations are indistinguishable for alanine, since all three substituents on the side-chain carbon atom C_β are the same.

For valine, however, the three staggered conformations are energetically different because the substituents on C_β are different; two of them are methyl groups and the other is a hydrogen atom (see Figure 1.8b–d). It is energetically most favorable to have the two methyl groups close to the small hydrogen atom bound to C_α, as shown in Figure 1.8b, and therefore this is the conformation most frequently found in proteins. An analysis of accurately determined protein structures has shown that most side chains have one or a few conformations that occur more frequently than the other possible staggered

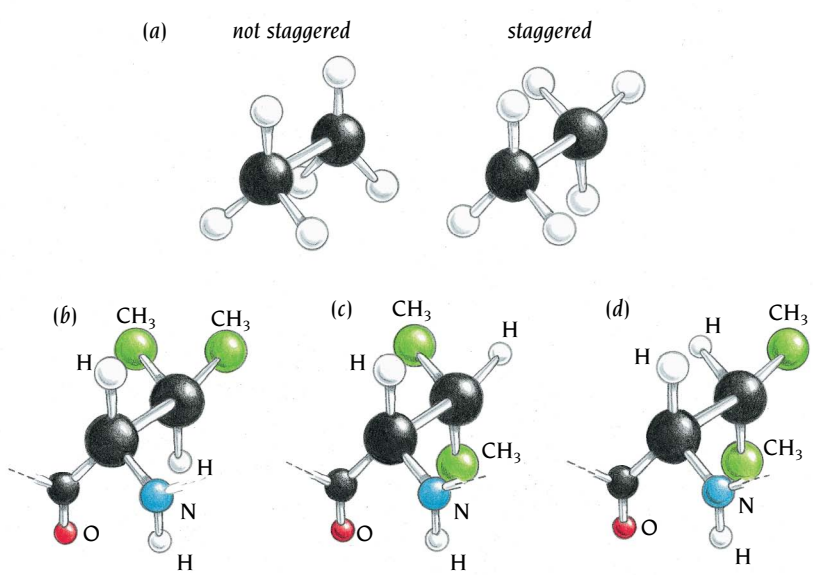


Figure 1.8 The staggered conformations are the most energetically favored conformations of two tetrahedrally coordinated carbon atoms. (a) A view along the C–C bond in ethane (CH_3CH_3) showing how the two carbon atoms can rotate so that their hydrogen atoms are either not staggered (aligned) or staggered.Three indistinguishable staggered conformations are obtained by a rotation of 120° around the C–C bond. (b–d) Similar views as in (a) of valine. The three staggered conformations are different for valine because the three groups attached to C_β are different. The first staggered conformation (b) is less crowded and energetically most favored because the two methyl groups bound to C_β are both close to the small H atom bound to C_α.

Petitioner Merck, Ex. 1014, p. 10

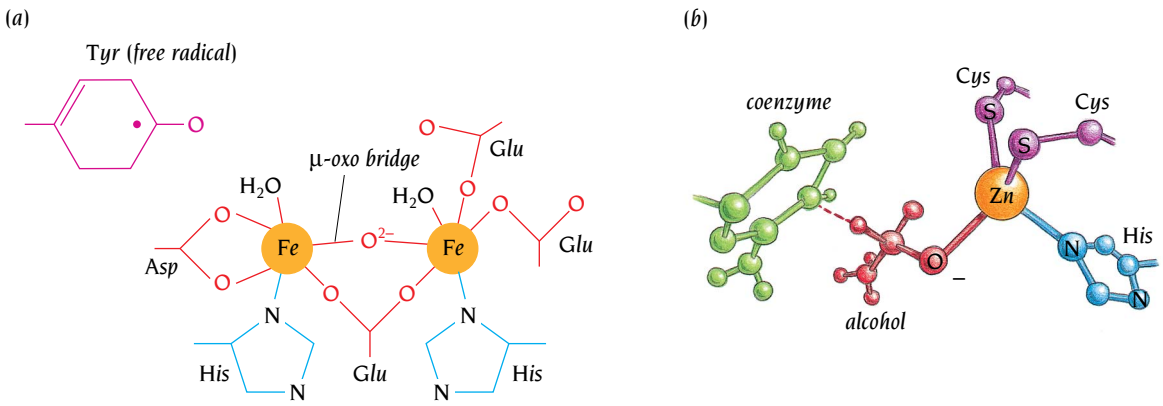


Figure 1.9 Examples of functionally important intrinsic metal atoms in proteins. (a) The di-iron center of the enzyme ribonucleotide reductase. Two iron atoms form a redox center that produces a free radical in a nearby tyrosine side chain. The iron atoms are bridged by a glutamic acid residue and a negatively charged oxygen atom called a μ-oxo bridge. The coordination of the iron atoms is completed by histidine, aspartic acid, and glutamic acid side chains as well as water molecules. (b) The catalytically active zinc atom in the enzyme alcohol dehydrogenase. The zinc atom is coordinated to the protein by one histidine and two cysteine side chains. During catalysis zinc binds an alcohol molecule in a suitable position for hydride transfer to the coenzyme moiety, a nicotinamide. [(a) Adapted from P. Nordlund et al., *Nature* 345: 593–598, 1990.]

conformations. These are called **rotamers**. Today, collections of these favored conformations, or rotamer libraries, are a standard tool in computer programs used for modeling protein structures.

Many proteins contain intrinsic metal atoms

The side chains of the 20 different amino acids listed in Panel 1.1 (pp. 6–7) have very different chemical properties and are utilized for a wide variety of biological functions. However, their chemical versatility is not unlimited, and for some functions metal atoms are more suitable and more efficient. Electron-transfer reactions are an important example. Fortunately the side chains of histidine, cysteine, aspartic acid, and glutamic acid are excellent metal ligands, and a fairly large number of proteins have recruited metal atoms as intrinsic parts of their structures; among the frequently used metals are iron, zinc, magnesium, and calcium. Several metallo proteins are discussed in detail in later chapters and it suffices here to mention briefly a few examples of iron and zinc proteins.

The most conspicuous use of iron in biological systems is in our blood, where the erythrocytes are filled with the oxygen-binding protein hemoglobin. The red color of blood is due to the iron atom bound to the heme group in hemoglobin. Similar heme-bound iron atoms are present in a number of proteins involved in electron-transfer reactions, notably cytochromes. A chemically more sophisticated use of iron is found in an enzyme, ribonucleotide reductase, that catalyzes the conversion of ribonucleotides to deoxyribonucleotides, an important step in the synthesis of the building blocks of DNA.

Ribonucleotide reductase from *Escherichia coli* and mammals contains a di-iron center (Figure 1.9a) that reacts with oxygen and oxidizes a nearby tyrosine side chain, producing a tyrosyl free radical that is essential for the catalysis. The two iron atoms in this iron center are close to each other and bridged by the oxygen atoms of a glutamic acid side chain as well as by an oxygen ion called a μ-oxo bridge. Glutamic acid, aspartic acid, and histidine side chains from the protein, as well as water molecules, complete the coordination sphere of the octahedrally coordinated iron atom.

Zinc is used to stabilize the DNA-binding regions of a class of transcription factors called zinc fingers, which are discussed in Chapter 10. Zinc ions also participate directly in catalytic reactions in many different enzymes by binding substrate molecules and providing a positive charge that influences the electronic arrangement of the substrate and thereby facilitates the catalytic reaction. One such example is found in the enzyme alcohol dehydrogenase, which in yeast produces alcohol during fermentation and in our livers detoxifies the alcohol we have consumed by oxidizing it. The enzyme provides a scaffold containing three zinc ligands—one histidine and two cysteine side chains—which sequester a zinc atom so that it binds alcohol as a fourth ligand in a tetrahedral coordination (Figure 1.9b).

Petitioner Merck, Ex. 1014, p. 11

Conclusion

All protein molecules are polymers built up from 20 different amino acids linked end-to-end by peptide bonds. The function of every protein molecule depends on its three-dimensional structure, which in turn is determined by its amino acid sequence, which in turn is determined by the nucleotide sequence of the structural gene.

Each amino acid has atoms in common, and these form the main chain of the protein. The remaining atoms form side chains that can be hydrophobic, polar, or charged.

The conformation of the whole main chain of a protein is determined by two conformational angles, phi (ϕ) and psi (ψ), for each amino acid. Only certain combinations of these angles are allowed because of steric hindrance between main-chain atoms and side-chain atoms, except for glycine.

Certain side-chain conformations are energetically more favorable than others. Computer programs used to model protein structures contain rotamer libraries of such favored conformations.

Many proteins contain intrinsic metal atoms that are functionally important. The most frequently used metals are iron, zinc, magnesium, and calcium. These metal atoms are mainly bound to the protein through the side chains of cysteine, histidine, aspartic acid, and glutamic acid residues.

Selected readings

Alberts, B., et al. *Molecular Biology of the Cell*, 3rd ed. New York: Garland, 1994.

Creighton, T.E. *Proteins: Structures and Molecular Properties*, 2nd ed. New York: Freeman, 1993.

Fletterick, R.J., Schroer, T., Matela, R.J. *Molecular Structure: Macromolecules in Three Dimensions*. Oxford, UK: Blackwell Scientific, 1985.

Judson, H.F. *The Eighth Day of Creation: Makers of the Revolution in Biology*. New York: Simon & Schuster, 1979.

Karlin, K.D. Metalloenzymes, structural motifs, and inorganic models. *Science* 261: 701–708, 1993.

Lesk, A. *Protein Architecture: A Practical Approach*. Oxford, UK: Oxford University Press, 1991.

Mathews, C.K., van Holde, K.E. *Biochemistry*. Menlo Park, CA: Benjamin/Cummings, 1990.

Perutz, M. *Protein Structure: New Approaches to Disease and Therapy*. New York: Freeman, 1992.

Petsko, G.A. On the other hand... *Science* 256: 1403–1404, 1992.

Ramachandran, G.N., Sasisekharan, V. Conformation of polypeptides and proteins. *Adv. Prot. Chem.* 28: 283–437, 1968.

Richardson, J.S., Richardson, D.C. Principles and patterns of protein conformation. In *Prediction of Protein Structure and the Principles of Protein Conformation* (ed. Fasman, G.D.), pp. 1–98. New York: Plenum, 1989.

Schulz, G.E., Schirmer, R.H. *Principles of Protein Structure*. New York: Springer, 1979.

Spiro, T.S. *Zinc Enzymes*. New York: Wiley, 1983.

Stryer, L. *Biochemistry*, 4th ed. New York: Freeman, 1995.

Petitioner Merck, Ex. 1014, p. 12

Motifs of Protein Structure

2

X-ray structural studies have played a major role in transforming chemistry from a descriptive science at the beginning of the twentieth century to one in which the properties of novel compounds can be predicted on theoretical grounds. When W.L. Bragg solved the very first crystal structure, that of rock salt, NaCl, the results completely changed prevalent concepts of bonding forces in ionic compounds.

The first x-ray crystallographic structural results on a globular protein molecule were reported for myoglobin (Figure 2.1) in 1958, and came as a shock to those who had hoped for simple, general principles of protein structure and function analogous to the simple and beautiful double-stranded DNA structure that had been determined five years before by James Watson and Francis Crick. John Kendrew at the Medical Research Council Laboratory of Molecular Biology, in Cambridge, UK, who determined the myoglobin structure to low resolution in 1958, expressed his disappointment about the complexity of the structure in the following words: "Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicted by any theory of protein structure."

In retrospect it is easy to see that such structural irregularity is actually required for proteins to fulfill their diverse functions. Information storage and transfer from DNA is essentially linear, and DNA molecules of very different information content can therefore have essentially the same gross structure. In contrast, proteins must recognize many thousands of different molecules in the cell by detailed three-dimensional interactions, which

Figure 2.1 Kendrew's model of the low-resolution structure of myoglobin shown in three different views. The sausage-shaped regions represent α helices, which are arranged in a seemingly irregular manner to form a compact globular molecule. (Courtesy of J.C. Kendrew.)

Petitioner Merck, Ex. 1014, p. 13

require diverse and irregular structures of the protein molecules. In spite of these requirements, there are regular features in protein structures, the most important of which is their **secondary structure.**

The interior of proteins is hydrophobic

When high-resolution studies of myoglobin became available, Kendrew noticed that the amino acids in the interior of the protein had almost exclusively hydrophobic side chains. This was one of the first important general principles to emerge from studies of protein structure. The main driving force for folding water-soluble globular protein molecules is to pack hydrophobic side chains into the interior of the molecule, thus creating a **hydrophobic core** and a hydrophilic surface.

The hydrophobic core is surprisingly densely packed with the side chains in the interior of the protein. Given the constraints of the different shapes of the hydrophobic side chains and considering that their positions must be compatible with the regular secondary structure in the interior of the protein, fitting these shapes into a densely packed core is like solving a three-dimensional jigsaw puzzle. In the few cases where there is a hole in the interior, the space is usually occupied by one or more water molecules that hydrogen-bond to internal polar groups. Such firmly bound internal water molecules can be regarded as integral parts of the protein structure.

There is a major problem, however, with creating such a hydrophobic core from a protein chain. To bring the side chains into the core, the main chain must also fold into the interior. The main chain is highly polar and therefore hydrophilic, with one hydrogen bond donor, NH, and one hydrogen bond acceptor, C′=O, for each peptide unit. In a hydrophobic environment, these main-chain polar groups must be neutralized by the formation of hydrogen bonds. This problem is solved in a very elegant way by the formation of regular secondary structure within the interior of the protein molecule. Such secondary structure is usually one of two types: **alpha helices** or **beta sheets.** Both types are characterized by hydrogen-bonding between the main-chain NH and C′=O groups, and they are formed when a number of consecutive residues have the same phi (ϕ), psi (ψ) angles.

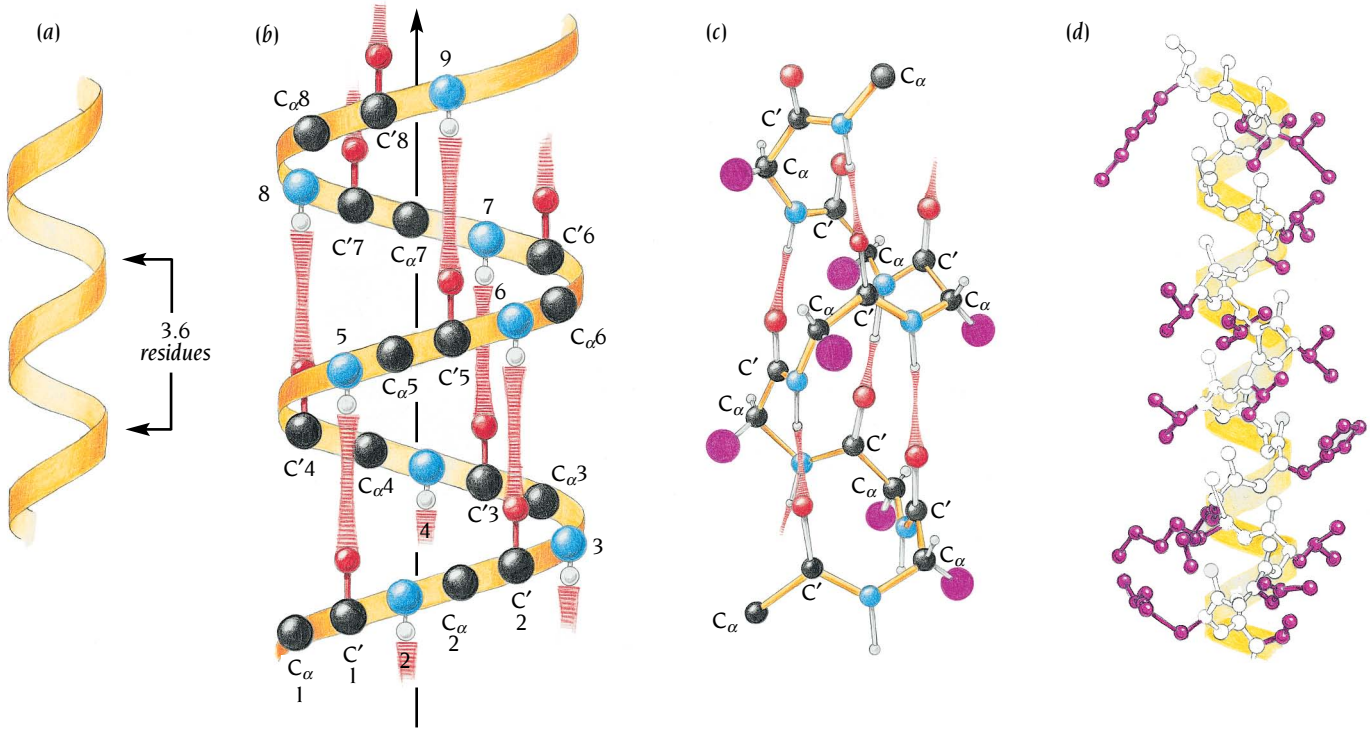
The secondary structure elements, formed in this way and held together by the hydrophobic core, provide a rigid and stable framework. They exhibit relatively little flexibility with respect to each other, and they are the best-defined parts of protein structures determined by both x-ray and NMR techniques. Functional groups of the protein are attached to this framework, either directly by their side chains or, more frequently, in loop regions that connect sequentially adjacent secondary structure elements. We will now have a closer look at these structural elements.

The alpha (α) helix is an important element of secondary structure

The α helix is the classic element of protein structure. It was first described in 1951 by Linus Pauling working at the California Institute of Technology. He predicted that it was a structure which would be stable and energetically favorable in proteins. He made this remarkable prediction on the basis of accurate geometrical parameters that he had derived for the peptide unit from the results of crystallographic analyses of the structures of a range of small molecules. This prediction almost immediately received strong experimental support from diffraction patterns obtained by Max Perutz in Cambridge, UK, from hemoglobin crystals and keratin fibers. It was completely verified from John Kendrew's high-resolution structure of myoglobin, where all secondary structure is helical.

Alpha helices in proteins are found when a stretch of consecutive residues all have the ϕ, ψ angle pair approximately –60° and –50°, corresponding to the allowed region in the bottom left quadrant of the

Petitioner Merck, Ex. 1014, p. 14



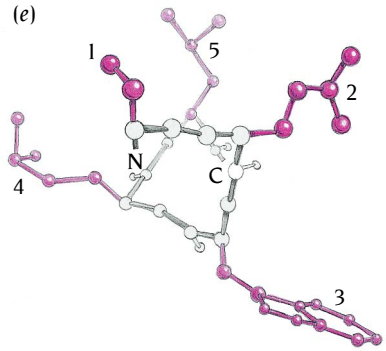


Figure 2.2 The α helix is one of the major elements of secondary structure in proteins. Main-chain N and O atoms are hydrogen-bonded to each other within α helices. (a) Idealized diagram of the path of the main chain in an α helix. Alpha helices are frequently illustrated in this way. There are 3.6 residues per turn in an α helix, which corresponds to 5.4 Å (1.5 Å per residue). (b) The same as (a) but with approximate positions for main-chain atoms and hydrogen bonds included. The arrow denotes the direction from the N-terminus to the C-terminus. (c) Schematic diagram of an α helix. Oxygen atoms are red, and N atoms are blue. Hydrogen bonds between O and N are red and striated. The side chains are represented as purple circles. (d) A ball-and-stick model of one α helix in myoglobin. The path of the main chain is outlined in yellow; side chains are purple. Main-chain atoms are not colored. (e) One turn of an α helix viewed down the helical axis. The purple side chains project out from the α helix.

Ramachandran plot (see Figure 1.7a). The α helix has 3.6 residues per turn with hydrogen bonds between C′=O of residue n and NH of residue $n + 4$ (Figure 2.2). Thus all NH and C′O groups are joined with hydrogen bonds except the first NH groups and the last C′O groups at the ends of the α helix. As a consequence, the ends of α helices are polar and are almost always at the surface of protein molecules.

Variations on the α helix in which the chain is either more loosely or more tightly coiled, with hydrogen bonds to residues $n + 5$ or $n + 3$ instead of $n + 4$ are called the π helix and 3_{10} helix, respectively. The 3_{10} helix has 3 residues per turn and contains 10 atoms between the hydrogen bond donor and acceptor, hence its name. Both the π helix and the 3_{10} helix occur rarely and usually only at the ends of α helices or as single-turn helices. They are not energetically favorable, since the backbone atoms are too tightly packed in the 3_{10} helix and so loosely packed in the π helix that there is a hole through the middle. Only in the α helix are the backbone atoms properly packed to provide a stable structure.

In globular proteins α helices vary considerably in length, ranging from four or five amino acids to over forty residues. The average length is around ten residues, corresponding to three turns. The rise per residue of an α helix is 1.5 Å along the helical axis, which corresponds to about 15 Å from one end to the other of an average α helix.

Petitioner Merck, Ex. 1014, p. 15

An α helix can in theory be either right-handed or left-handed depending on the screw direction of the chain. A left-handed α helix is not, however, allowed for L-amino acids due to the close approach of the side chains and the C′O group. Thus the α helix that is observed in proteins is almost always right-handed. Short regions of left-handed α helices (3–5 residues) occur only occasionally.

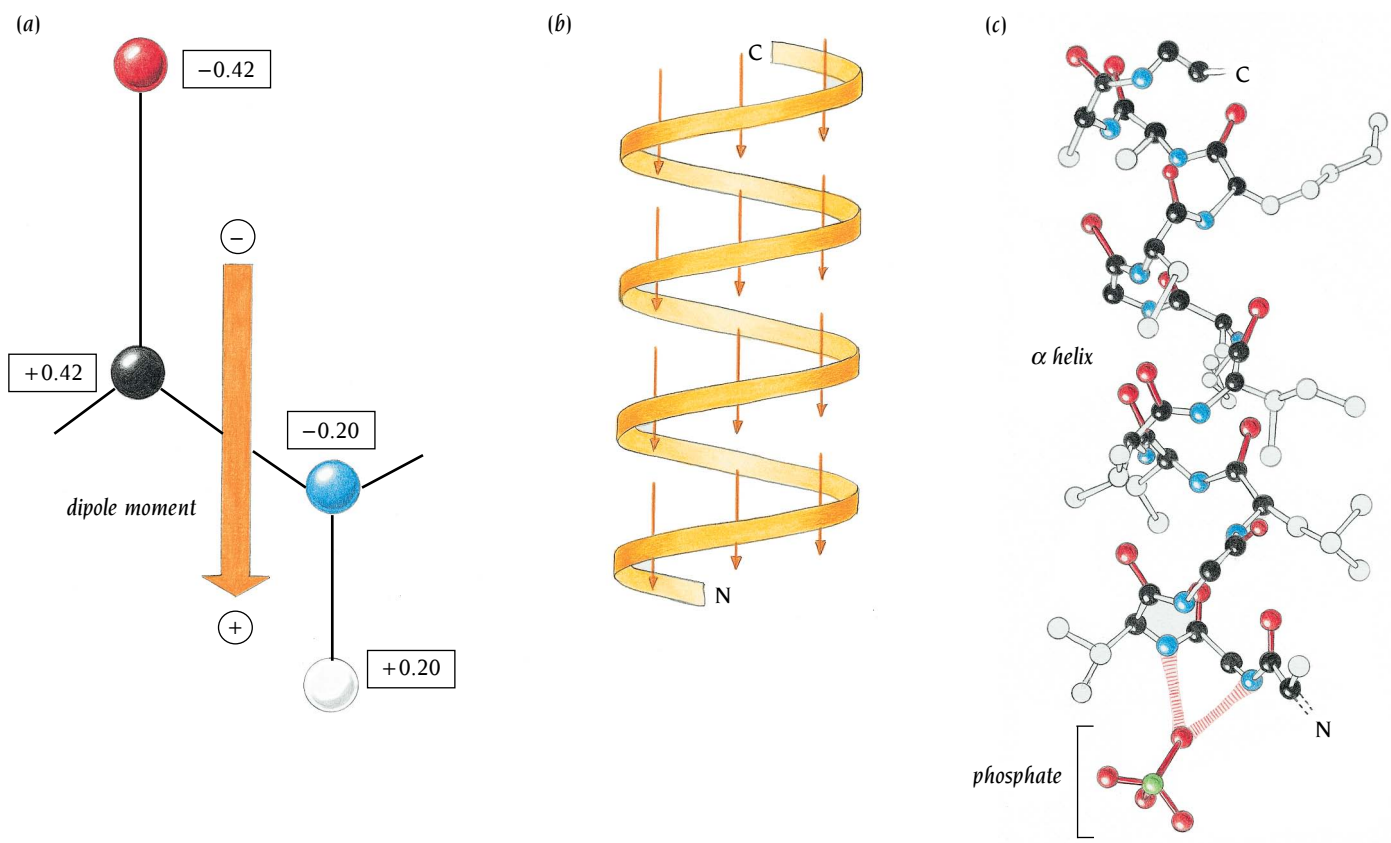
The α helix has a dipole moment

All the hydrogen bonds in an α helix point in the same direction because the peptide units are aligned in the same orientation along the helical axis. Since a peptide unit has a dipole moment arising from the different polarity of NH and C′O groups, these dipole moments are also aligned along the helical axis (Figure 2.3). The overall effect is a significant net dipole for the α helix that gives a partial positive charge at the amino end and a partial negative charge at the carboxy end of the α helix. The magnitude of this dipole moment corresponds to about 0.5–0.7 unit charge at each end of the helix. These charges would be expected to attract ligands of opposite charge and negatively charged ligands, especially when they contain phosphate groups and frequently bind at the N-termini of α helices. In contrast, positively charged ligands rarely bind at the C-terminus. This may be because, in addition to the dipole effect, the N-terminus of an α helix has free NH groups with favorable geometry to position phosphate groups by specific hydrogen bonds (see Figure 2.3). Such ligand-binding occurs frequently in proteins; it provides examples of specific binding through main-chain conformation in which side chains are not involved.

Some amino acids are preferred in α helices

The amino acid side chains project out from the α helix (see Figure 2.2e) and do not interfere with it, except for proline. The last atom of the proline side

Figure 2.3 Negatively charged groups such as phosphate ions frequently bind to the amino ends of α helices. The dipole moment of an α helix as well as the possibility of hydrogen-bonding to free NH groups at the end of the helix favors such binding. (a) The dipole of a peptide unit. Values in boxes give the approximate fractional charges of the atoms of the peptide unit. (b) The dipoles of peptide units are aligned along the α-helical axis, which creates an overall dipole moment of the α helix, positive at the amino end and negative at the carboxy end. (c) A phosphate group hydrogen-bonded to the NH end of an α helix. Nitrogen atoms are blue; oxygen atoms are red; main-chain carbon atoms are black; and phosphorus is green.



Petitioner Merck, Ex. 1014, p. 16

Table 2.1 Amino acid sequences of three α helices

1.	Leu	Ser	Phe	Ala	Ala	Ala	Met	Asn	Gly	Leu	Ala
2.	Ile	Asn	Glu	Gly	Phe	Asp	Leu	Leu	Arg	Ser	Gly
3.	Lys	Glu	Asp	Ala	Lys	Gly	Lys	Ser	Glu	Glu	Glu

The first sequence is from the enzyme citrate synthase, residues 260–270, which form a buried helix; the second sequence is from the enzyme alcohol dehydrogenase, residues 355–365, which form a partially exposed helix; and the third sequence is from troponin-C, residues 87–97, which form a completely exposed helix. Charged residues are colored red, polar residues are blue, and hydrophobic residues are green.

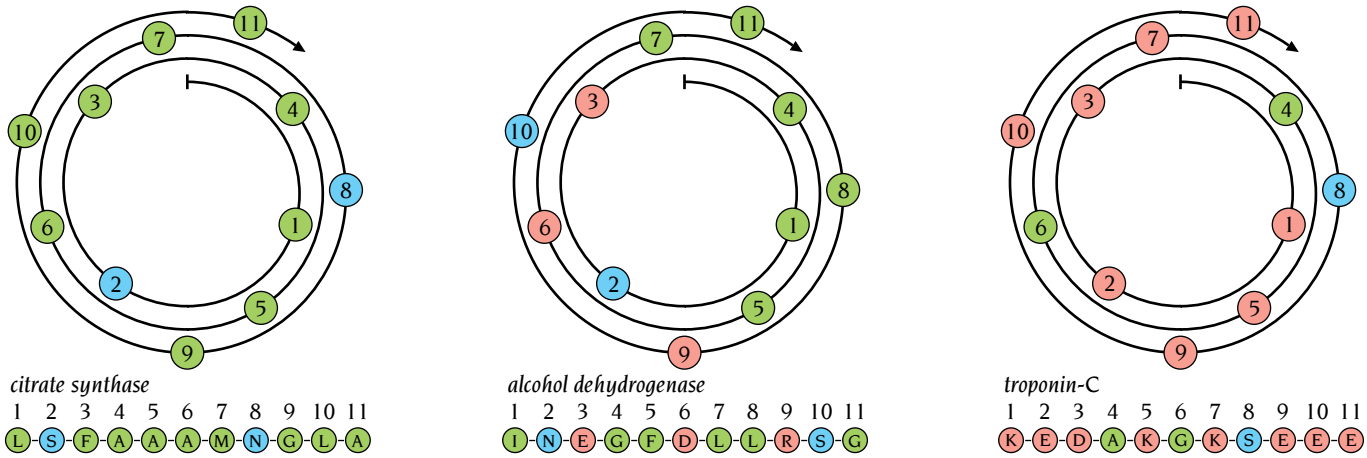
chain is bonded to the main-chain N atom, which forms a ring structure, C_α–CH_2–CH_2–CH_2–N (see Panel 1.1, p. 7). This prevents the N atom from participating in hydrogen-bonding and also provides some steric hindrance to the α-helical conformation. Proline fits very well in the first turn of an α helix, but it usually produces a significant bend if it is anywhere else in the helix. Such bends occur in many α helices, not just in those few that contain a proline in the middle. Therefore, although we can predict that a proline residue may cause a bend in an α helix, it does not follow that all bends result from the presence of proline.

Different side chains have been found to have weak but definite preferences either for or against being in α helices. Thus Ala (A), Glu (E), Leu (L), and Met (M) are good α-helix formers, while Pro (P), Gly (G), Tyr (Y), and Ser (S) are very poor. Such preferences were central to all early attempts to predict secondary structure from amino acid sequence, but they are not strong enough to give accurate predictions.

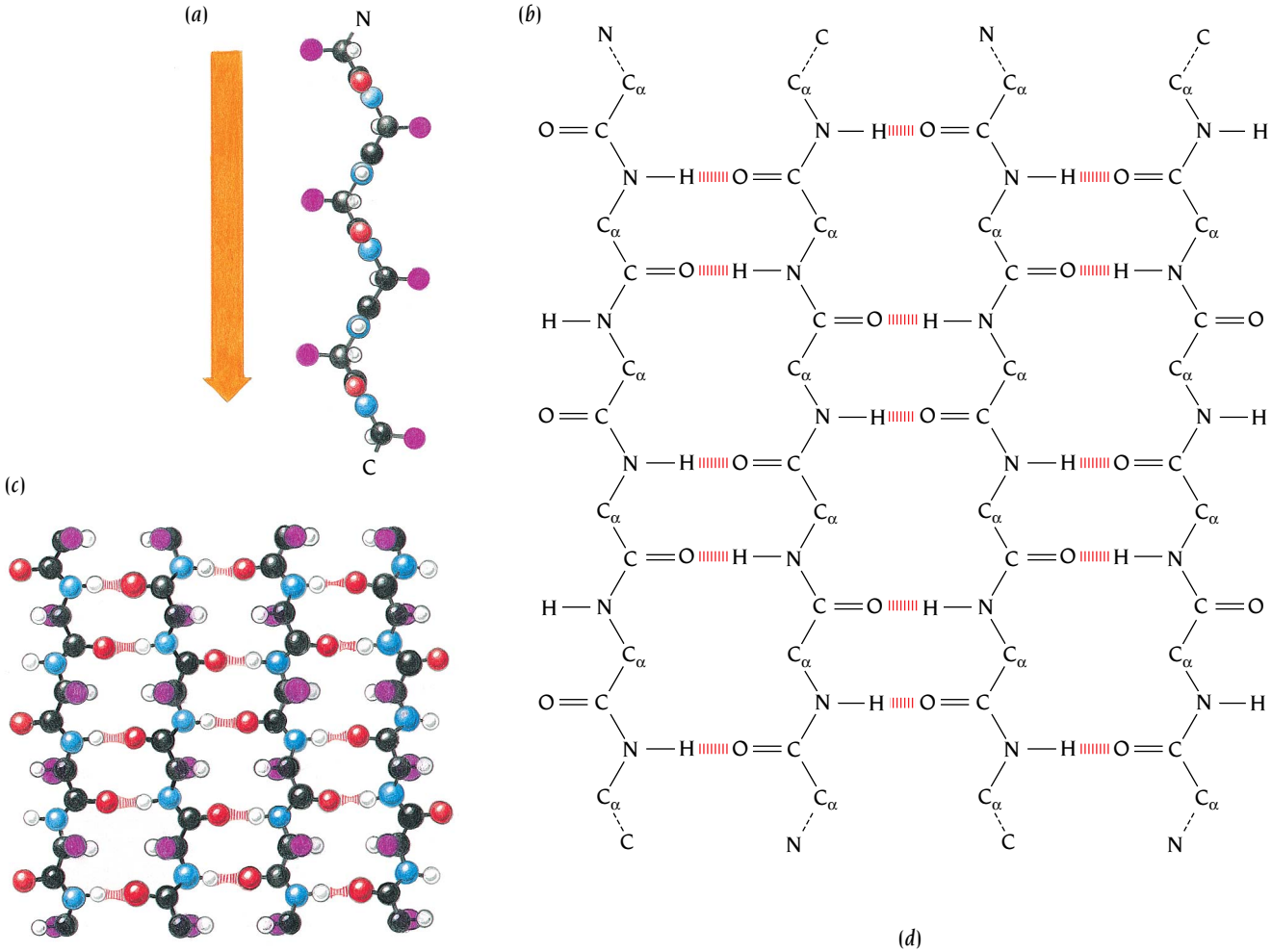
The most common location for an α helix in a protein structure is along the outside of the protein, with one side of the helix facing the solution and the other side facing the hydrophobic interior of the protein. Therefore, with 3.6 residues per turn, there is a tendency for side chains to change from hydrophobic to hydrophilic with a periodicity of three to four residues. Although this trend can sometimes be seen in the amino acid sequence, it is not strong enough for reliable structural prediction by itself, because residues that face the solution can be hydrophobic and, furthermore, α helices can be either completely buried within the protein or completely exposed. Table 2.1 shows examples of the amino acid sequences of a totally buried, a partially buried, and a completely exposed α helix.

A convenient way to illustrate the amino acid sequences in helices is the **helical wheel** or spiral. Since one turn in an α helix is 3.6 residues long, each residue can be plotted every 360/3.6 = 100° around a circle or a spiral, as shown in Figure 2.4. Such a plot shows the projection of the position of the

Figure 2.4 The helical wheel or spiral. Amino acid residues are plotted every 100° around the spiral, following the sequences given in Table 2.1. The following color code is used: green is an amino acid with a hydrophobic side chain, blue is a polar side chain, and red is a charged side chain. The first helix is all hydrophobic, the second is polar on one side and hydrophobic on the other side, and the third helix is all polar.



Petitioner Merck, Ex. 1014, p. 17



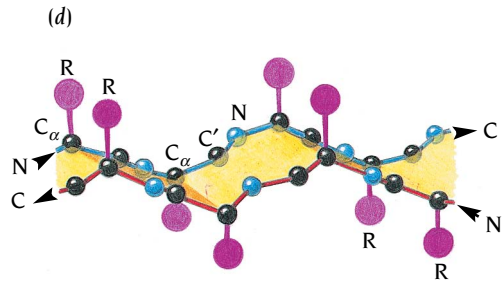


Figure 2.5 Schematic illustrations of antiparallel β sheets. Beta sheets are the second major element of secondary structure in proteins. The β strands are either all antiparallel as in this figure or all parallel or mixed as illustrated in following figures. (a) The extended conformation of a β strand. Side chains are shown as purple circles. The orientation of the β strand is at right angles to those of (b) and (c). A β strand is schematically illustrated as an arrow, from N to C terminus. (b) Schematic illustration of the hydrogen bond pattern in an antiparallel β sheet. Main-chain NH and O atoms within a β sheet are hydrogen bonded to each other. (c) A ball-and-stick version of (b). Oxygen atoms are red; nitrogen atoms are blue. The hydrogen atom in N–H···O is white. The carbon atom in the main chain, C_α, is black. Side chains are illustrated by one purple atom. The orientation of the β strands is different from that in (a). (d) Illustration of the pleat of a β sheet. Two antiparallel β strands are viewed from the side of the β sheet. Note that the directions of the side chains, R (purple), follow the pleat, which is emphasized in yellow.

residues onto a plane perpendicular to the helical axis. Residues on one side of the helix are plotted on one side of the spiral. The three helices whose sequences are given in Table 2.1 are plotted in this way in Figure 2.4, using a color code for hydrophobic, polar, and charged residues. It is immediately obvious that one side of the helix from alcohol dehydrogenase is hydrophilic and the other side hydrophobic.

Alpha helices that cross membranes are in a hydrophobic environment. Therefore, most of their side chains are hydrophobic. Long regions of hydrophobic residues in the amino acid sequence of a protein that is membrane-bound can therefore be predicted with a high degree of confidence to be transmembrane helices, as will be discussed in Chapter 12.

Petitioner Merck, Ex. 1014, p. 18

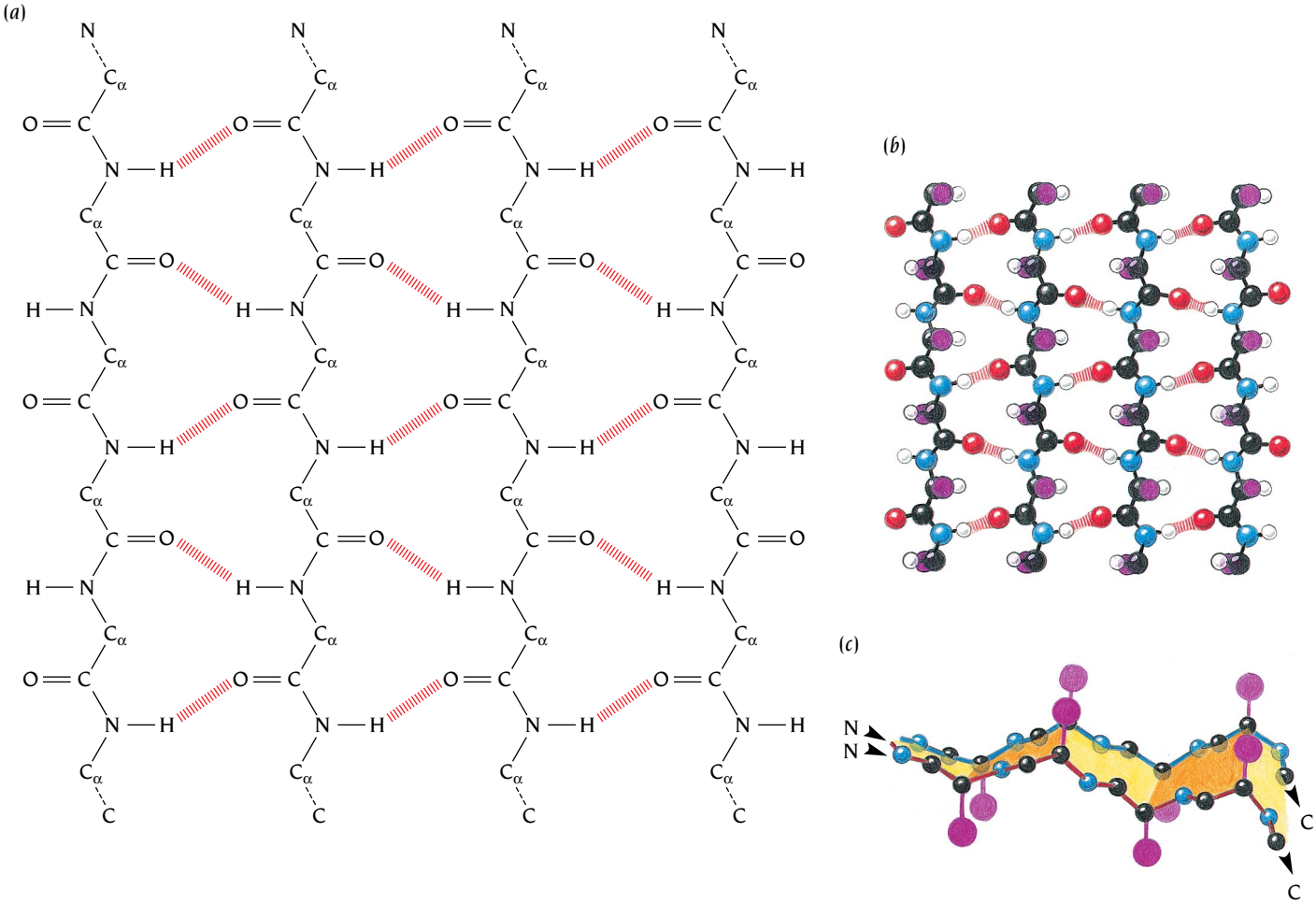


Figure 2.6 Parallel β sheet. (a) Schematic diagram showing the hydrogen bond pattern in a parallel β sheet. (b) Ball-and-stick version of (a). The same color scheme is used as in Figure 2.5c. (c) Schematic diagram illustrating the pleat of a parallel β sheet.

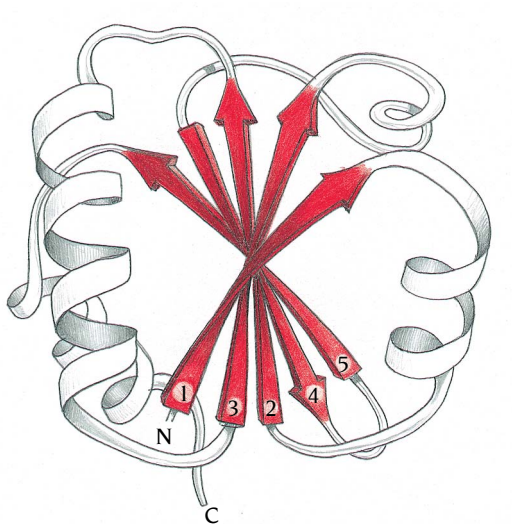
Beta (β) sheets usually have their β strands either parallel or antiparallel

The second major structural element found in globular proteins is the β sheet. This structure is built up from a combination of several regions of the polypeptide chain, in contrast to the α helix, which is built up from one continuous region. These regions, β strands, are usually from 5 to 10 residues long and are in an almost fully extended conformation with ϕ, ψ angles within the broad structurally allowed region in the upper left quadrant of the Ramachandran plot (see Figure 1.7). These β strands are aligned adjacent to each other (see Figures 2.5 and 2.6) such that hydrogen bonds can form between C′=O groups of one β strand and NH groups on an adjacent β strand and vice versa. The β sheets that are formed from several such β strands are "**pleated**" with C_α atoms successively a little above and below the plane of the β sheet. The side chains follow this pattern such that within a β strand they also point alternately above and below the β sheet.

Beta strands can interact in two ways to form a pleated sheet. Either the amino acids in the aligned β strands can all run in the same biochemical direction, amino terminal to carboxy terminal, in which case the sheet is described as **parallel**, or the amino acids in successive strands can have alternating directions, amino terminal to carboxy terminal followed by carboxy terminal to amino terminal, followed by amino terminal to carboxy terminal, and so on, in which case the sheet is called **antiparallel.** Each of the two forms has a distinctive pattern of hydrogen-bonding. The antiparallel β sheet (Figure 2.5) has narrowly spaced hydrogen bond pairs that alternate with widely spaced pairs. Parallel β sheets (Figure 2.6) have evenly spaced hydrogen bonds that bridge the β strands at an angle. Within both types of β sheets all possible main-chain hydrogen bonds are formed, except for the two flanking strands of the β sheet that have only one neighboring β strand.

Petitioner Merck, Ex. 1014, p. 19

(*a*)



(*b*)

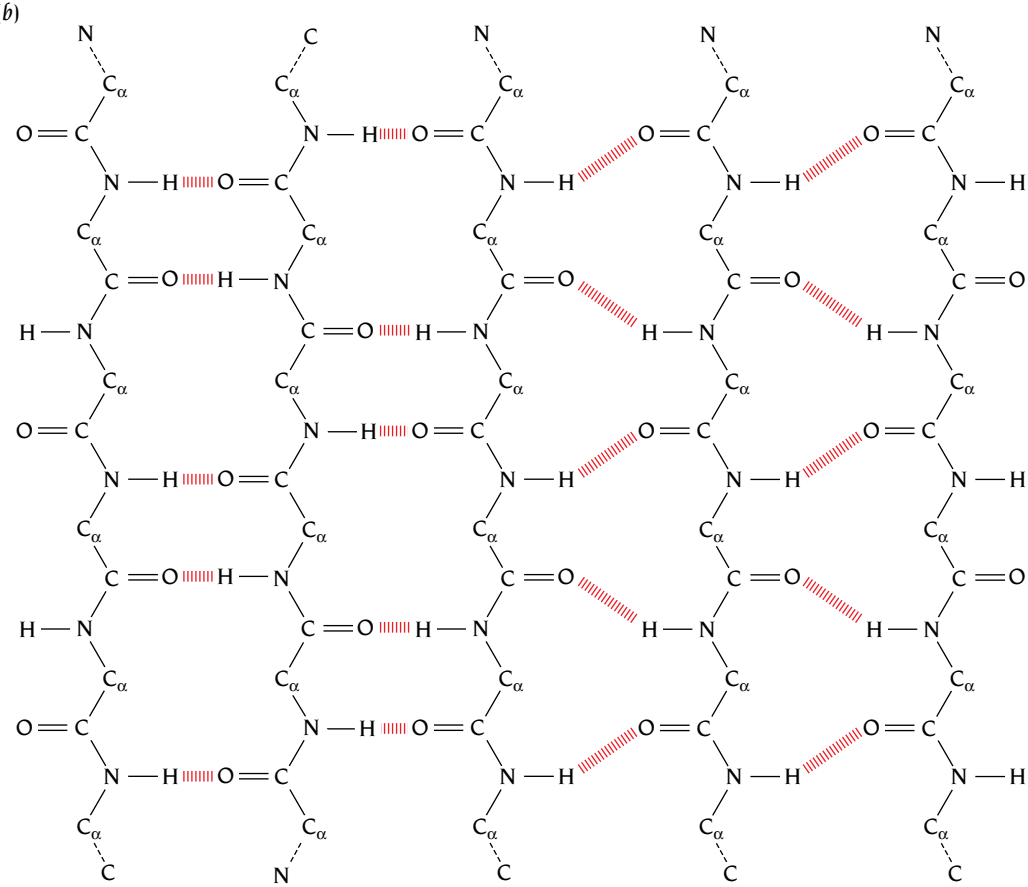


Figure 2.7 (a) Illustration of the twist of β sheets. Beta strands are drawn as arrows from the amino end to the carboxy end of the β strand in this schematic drawing of the protein thioredoxin from *E. coli,* the structure of which was determined in the laboratory of Carl Branden, Uppsala, Sweden, to 2.8 Å resolution. The mixed β sheet is viewed from one of its ends. (b) The hydrogen bonds between the β strands in the mixed β sheet of the same protein. [(a) Adapted from B. Furugren.]

Beta strands can also combine into mixed β sheets with some β strand pairs parallel and some antiparallel. There is a strong bias against mixed β sheets; only about 20% of the strands inside the β sheets of known protein structures have parallel bonding on one side and antiparallel bonding on the other. Figure 2.7 illustrates how the hydrogen bonds between the β strands are arranged in a mixed β sheet.

As they occur in known protein structures, almost all β sheets—parallel, antiparallel, and mixed—have twisted strands. This twist always has the same handedness as that shown in Figure 2.7, which is defined as a right-handed twist.

Petitioner Merck, Ex. 1014, p. 20

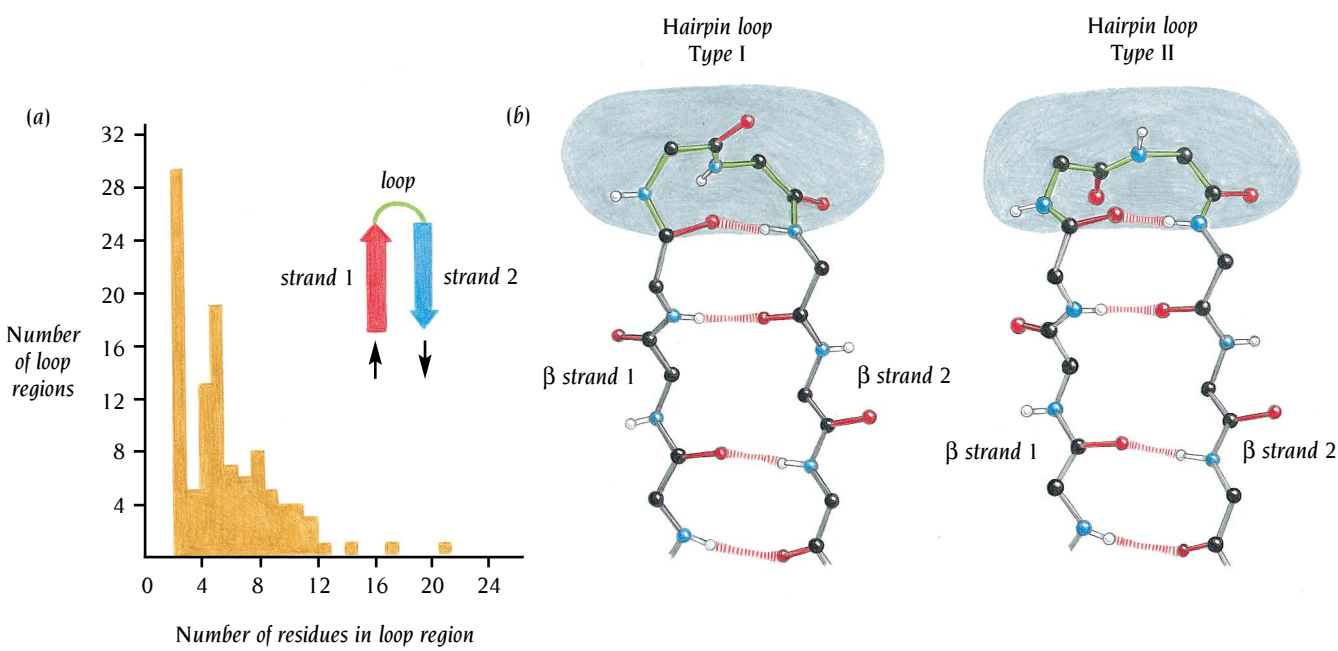
Loop regions are at the surface of protein molecules

Most protein structures are built up from combinations of secondary structure elements, α helices and β strands, which are connected by **loop regions** of various lengths and irregular shape. A combination of secondary structure elements forms the stable hydrophobic core of the molecule. The loop regions are at the surface of the molecule. The main-chain C′=O and NH groups of these loop regions, which in general do not form hydrogen bonds to each other, are exposed to the solvent and can form hydrogen bonds to water molecules.

Loop regions exposed to solvent are rich in charged and polar hydrophilic residues. This has been used in several prediction schemes, and it has proved possible to predict loop regions from an amino acid sequence with a higher degree of confidence than α helices or β strands, which is ironic since the loops have irregular structures.

When homologous amino acid sequences from different species are compared, it is found that insertions and deletions of a few residues occur almost exclusively in the loop regions. During evolution, cores are much more stable than loops. Intron positions are also often found at sites in structural genes that correspond to loop regions in the protein structure. Since proteins that exhibit sequence homology in general have similar core structures, it is apparent that the specific arrangement of secondary structure elements in the core is rather insensitive to the lengths of the loop regions. In addition to their function as connecting units between secondary structure elements, loop regions frequently participate in forming binding sites and enzyme active sites. Thus antigen-binding sites in antibodies are built up from six loop regions, which vary both in length and in amino acid sequence between different antibodies. Modeling antigen-binding sites from a known antibody sequence (discussed in Chapter 17) is thus essentially a problem of modeling the three-dimensional structures of loop regions since the core structures of all antibodies are very similar. Such model building has been facilitated by the recent findings that loop regions have preferred structures. Surveys of known three-dimensional structures of loops have shown that they fall into a rather limited set of structures and are not a random collection of possible structures (Figure 2.8). Loop regions that connect two adjacent antiparallel β strands are called **hairpin loops.** Short hairpin loops are usually called **reverse turns** or simply turns. Figure 2.8b shows two of the most frequently

Figure 2.8 Adjacent antiparallel β strands are joined by hairpin loops. Such loops are frequently short and do not have regular secondary structure. Nevertheless, many loop regions in different proteins have similar structures. (a) Histogram showing the frequency of hairpin loops of different lengths in 62 different proteins. (b) The two most frequently occurring two-residue hairpin loops; Type I turn to the left and Type II turn to the right. Bonds within the hairpin loop are green. [(a) Adapted from B.L. Sibanda and J.M. Thornton, *Nature* 316: 170–174, 1985.]



Petitioner Merck, Ex. 1014, p. 21

occurring turns; the Type I turn and the Type II turn. The Type II turn usually has a glycine residue as the second of the two residues in the turn.

Long loop regions are often flexible and can frequently adopt several different conformations, making them "invisible" in x-ray structure determinations and undetermined in NMR studies. Such loops are frequently involved in the function of the protein and can switch from an "open" conformation, which allows access to the active site, to a "closed" conformation, which shields reactive groups in the active site from water.

Long loops are in many cases susceptible to proteolytic degradation. One specific type of long loop, the omega loop, is compact with good internal packing interactions and is therefore quite stable. Other long loops, which by themselves would be attacked by proteolytic enzymes, are stabilized and protected by binding metal ions, especially calcium.

Schematic pictures of proteins highlight secondary structure

All pictorial representations of molecules are simplified versions of our current model of real molecules, which are quantum mechanical, probabilistic collections of atoms as both particles and waves. These are difficult to illustrate. Therefore we use different types of simplified representations, including space-filling models; ball-and-stick models, where atoms are spheres and bonds are sticks; and models that illustrate surface properties. The most detailed representation is the ball-and-stick model. However, a model of a protein structure where all atoms are displayed is confusing because of the sheer amount of information present (Figure 2.9a).

A two-dimensional picture of such a model is impossible to interpret. Even if the side-chain atoms are stripped off, it is still difficult to extract

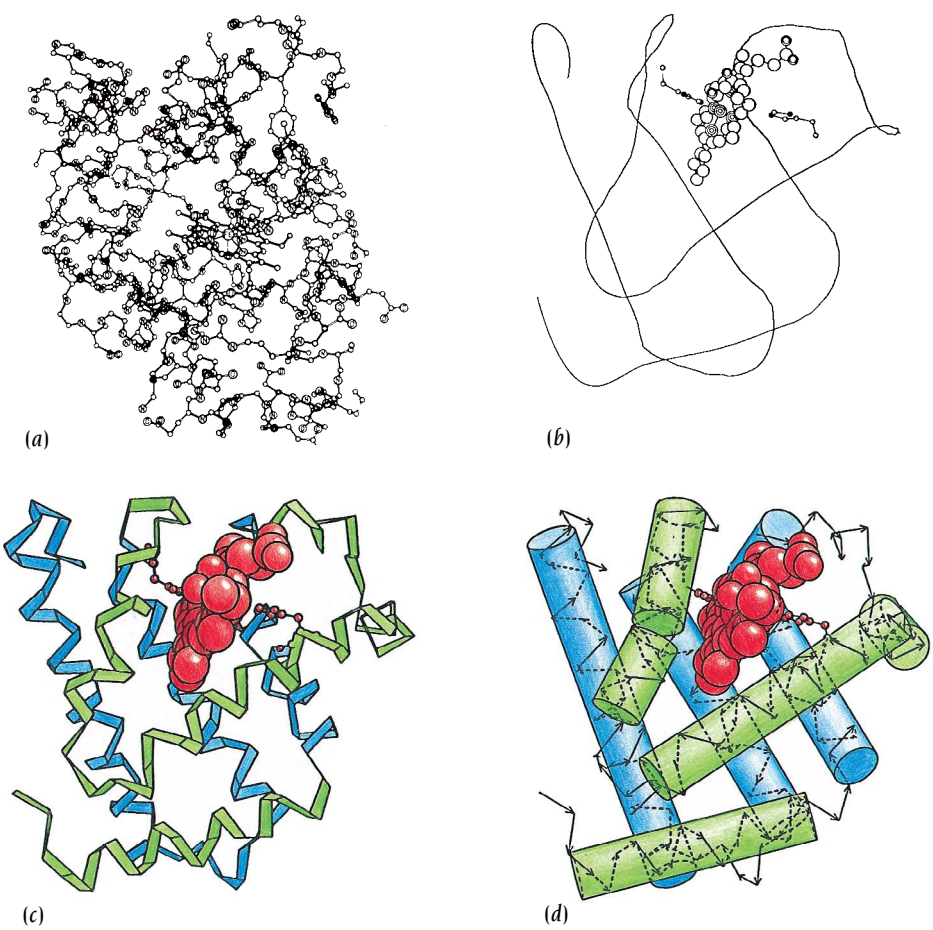


Figure 2.9 (a) The structure of myoglobin displaying all atoms as small circles connected by straight lines. Even though only side chains at the surface of the molecule are shown, the picture contains so many atoms that such a two-dimensional representation is very confusing and very little information can be gained from it. (b–d) Computer-generated schematic diagrams at different degrees of simplification of the structure of myoglobin. [(a) Half of a stereo diagram by H.C. Watson, *Prog. Stereochem.* 4: 299–333, 1969, by permission of Plenum Press. (b–d) From Arthur Lesk, Protein Architecture: A Practical Approach, 1991, Oxford University Press.]

Petitioner Merck, Ex. 1014, p. 22

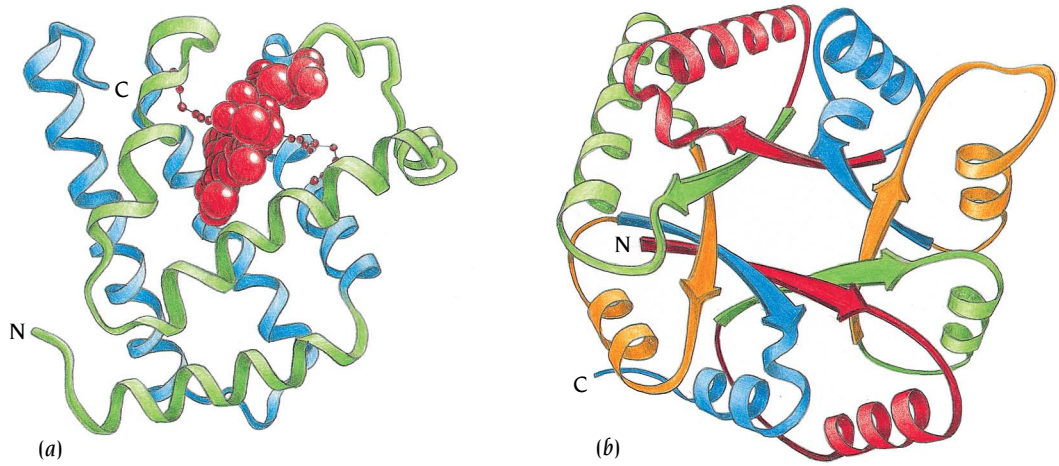


Figure 2.10 Examples of schematic diagrams of the type pioneered by Jane Richardson. Diagram (a) illustrates the structure of myoglobin in the same orientation as the computer-drawn diagrams of Figures 2.9b–d. Diagram (b), which is adapted from J. Richardson, illustrates the structure of the enzyme triosephosphate isomerase, determined to 2.5 Å resolution in the laboratory of David Phillips, Oxford University. Such diagrams can easily be obtained from databases of protein structures, such as PDB, SCOP or CATH, available on the World Wide Web.

meaningful information from a flat picture of such a model, mainly because it is difficult to see the relationships between the secondary structural elements. Since these elements dominate the structure, the picture becomes clearer if they are simplified and highlighted in some way. This is usually done by representing the path of the polypeptide chain by three different symbols: cylinders for α helices; arrows for β strands, which give the direction of the strands from amino to carboxy end; and ribbons for the remaining parts. Such schematic diagrams give good and very useful overall views of protein structures but, of course, give no detailed information. Details are best studied on graphic display systems where one can manipulate the computer-generated model on the screen; in this way the power of the graphics device makes it possible for the viewer to study much greater detail intelligibly.

The Kinemage Supplement, produced by Jane Richardson at Duke University, is available from the publisher as a complement to this book for readers with access to a personal computer. The system is easy to use and provides interactive three-dimensional viewing of many of the structures discussed in various chapters.

Jane Richardson has also made a very popular collection of **schematic diagrams** of various protein molecules with an artistic touch that gives an aesthetic impression without losing too much accuracy. Arthur Lesk at the MRC Laboratory of Molecular Biology in Cambridge, UK, and Karl Hardman at IBM pioneered the use of computer programs to generate schematic diagrams on a computer display from a list of atomic coordinates of the main-chain atoms. Figures 2.9b–d and 2.10 show representative examples of both Lesk-type and Richardson-type diagrams.

Topology diagrams are useful for classification of protein structures

It is very convenient for some purposes to have an even more simplified schematic representation of the secondary structure elements, especially for β sheets. The most characteristic features of a β sheet are the number of strands, their relative directions (parallel or antiparallel), and how the strands are connected along the polypeptide chain (the strand order). This information can be easily conveyed through simple diagrams of connected arrows like those in Figure 2.11, where such simple **topology diagrams** are compared to the more elaborate Richardson diagrams. The twist of the β sheet is not represented in these topology diagrams. They are, nevertheless, very helpful when used to compare β structures and to analyze and present data in computerized database searches of similar structures. Such topology diagrams will be used frequently in this book.

Petitioner Merck, Ex. 1014, p. 23

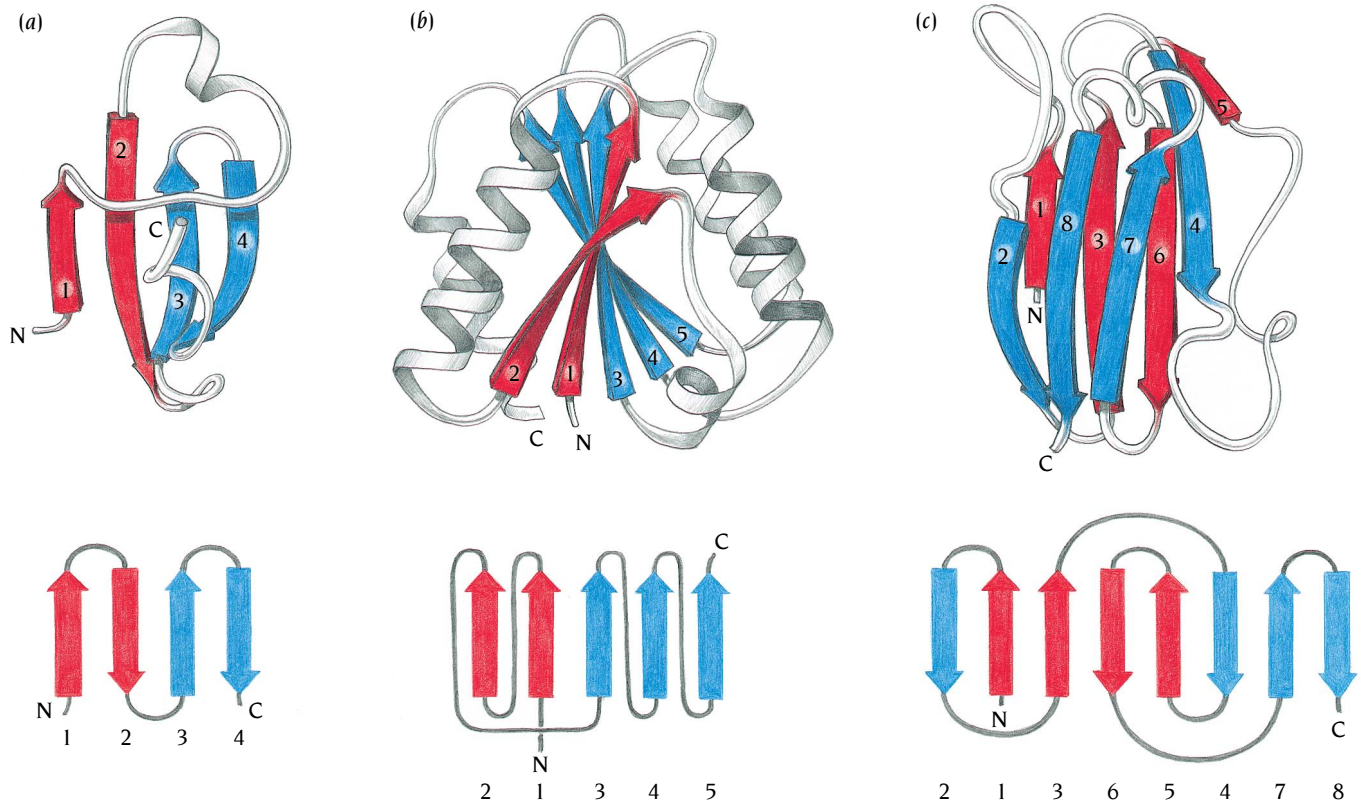


Figure 2.11 Beta sheets are usually represented simply by arrows in topology diagrams that show both the direction of each β strand and the way the strands are connected to each other along the polypeptide chain. Such topology diagrams are here compared with more elaborate schematic diagrams for different types of β sheets. (a) Four strands. Antiparallel β sheet in one domain of the enzyme aspartate transcarbamoylase. The structure of this enzyme has been determined to 2.8 Å resolution in the laboratory of William Lipscomb, Harvard University. (b) Five strands. Parallel β sheet in the redox protein flavodoxin, the structure of which has been determined to 1.8 Å resolution in the laboratory of Martha Ludwig, University of Michigan. (c) Eight strands. Antiparallel barrel in the electron carrier plastocyanin. This is a closed barrel where the sheet is folded such that β strands 2 and 8 are adjacent. The structure has been determined to 1.6 Å resolution in the laboratory of Hans Freeman in Sydney, Australia. (Adapted from J. Richardson.)

Secondary structure elements are connected to form simple motifs

Simple combinations of a few secondary structure elements with a specific geometric arrangement have been found to occur frequently in protein structures. These units have been called either supersecondary structures or **motifs.** We will use the term "motif" throughout this book. Some of these motifs can be associated with a particular function such as DNA binding; others have no specific biological function alone but are part of larger structural and functional assemblies.

The simplest motif with a specific function consists of two α helices joined by a loop region. Two such motifs, each with its own characteristic geometry and amino acid sequence requirements, have been observed as parts of many protein structures (Figure 2.12).

One of these motifs, called the helix-turn-helix motif, is specific for DNA binding and is described in detail in Chapters 8 and 9. The second motif is specific for calcium binding and is present in parvalbumin, calmodulin, troponin-C, and other proteins that bind calcium and thereby regulate cellular activities. This calcium-binding motif was first found in 1973 by Robert Kretsinger, University of Virginia, when he determined the structure of parvalbumin to 1.8 Å resolution.

Parvalbumin is a muscle protein with a single polypeptide chain of 109 amino acids. Its function is uncertain, but calcium binding to this protein probably plays a role in muscle relaxation. The helix-loop-helix motif appears three times in this structure, in two of the cases there is a calcium-binding site. Figure 2.13 shows this motif which is called an **EF hand** because the fifth and sixth helices from the amino terminus in the structure of parvalbumin, which were labeled E and F, are the parts of the structure that were originally used to illustrate calcium binding by this motif. Despite this trivial origin, the name has remained in the literature.

Petitioner Merck, Ex. 1014, p. 24

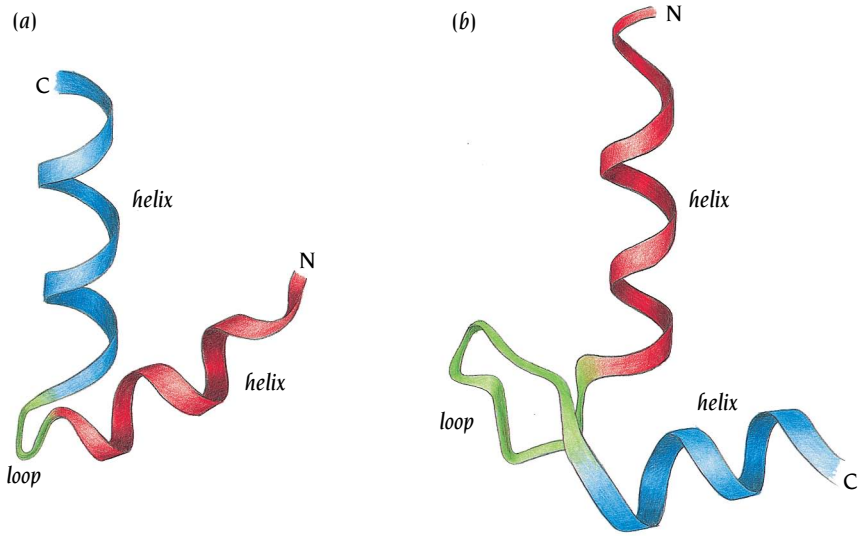
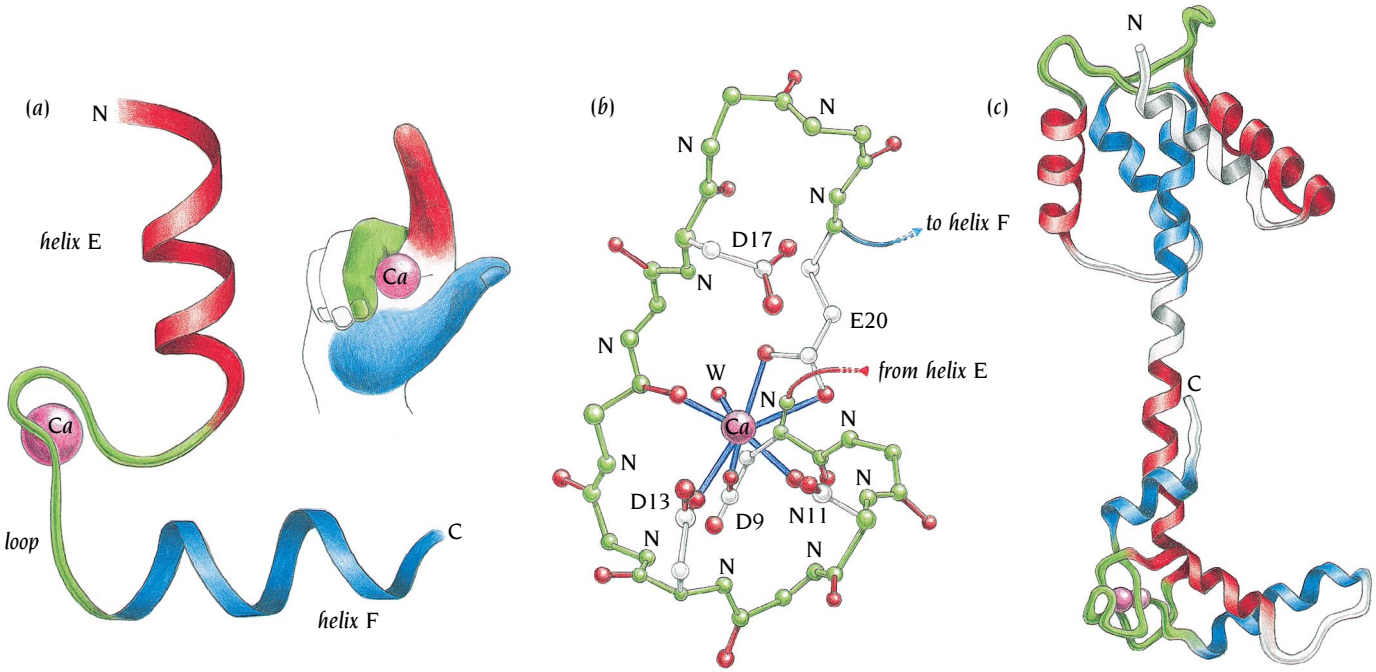


Figure 2.12 Two α helices that are connected by a short loop region in a specific geometric arrangement constitute a helix-turn-helix motif. Two such motifs are shown: the DNA-binding motif (a), which is further discussed in Chapter 8, and the calcium-binding motif (b), which is present in many proteins whose function is regulated by calcium.

The loop region between the two α helices binds the calcium atom. Carboxyl side chains from Asp and Glu, main-chain C′=O and H_2O form the ligands to the metal atom (see Figure 2.13b). Thus both the specific main-chain conformation of the loop and specific side chains are required to provide the function of this motif. The helix-loop-helix motif provides a scaffold that holds the calcium ligands in the proper position to bind and release calcium.

When Kretsinger analyzed in detail the structure of the calcium-binding motif in parvalbumin in 1973, he deduced a set of constraints that an amino acid sequence must conform to in order to form such a motif, one of the first to be recognized in protein structures. The motif comprises two α helices, E and F, that flank a loop of 12 contiguous residues. Five of the loop residues are calcium ligands, and their side chains should contain an oxygen atom and preferably be Asp or Glu (Table 2.2). Residue 6 of the loop must be a glycine because the side chain of any other residue would disturb the structure of the motif. Finally, a number of side chains form a hydrophobic core between the α helices and thus must be hydrophobic. Applying these constraints, Kretsinger predicted that several different calcium-binding proteins could form this motif. Among these proteins were the important muscle

Figure 2.13 Schematic diagrams of the calcium-binding motif. (a) The calcium-binding motif is symbolized by a right hand. Helix E (red) runs from the tip to the base of the forefinger. The flexed middle finger corresponds to the green loop region of 12 residues that binds calcium (pink). Helix F (blue) runs to the end of the thumb. (b) The calcium atom is bound to one of the motifs in the muscle protein troponin-C through six oxygen atoms: one each from the side chains of Asp (D) 9, Asn (N) 11, and Asp (D) 13; one from the main chain of residue 15; and two from the side chain of Glu (E) 20. In addition, a water molecule (W) is bound to the calcium atom. (c) Schematic diagram illustrating that the structure of troponin-C is built up from four EF motifs—colored as in (a). Two of these bind Ca (pink balls) in the molecules that were used for the structure determination. (Adapted from a diagram by J. Richardson in O. Herzberg and M. James, *Nature* 313: 653–659, 1985.)



Petitioner Merck, Ex. 1014, p. 25

Table 2.2 Amino acid sequences of calcium-binding EF motifs in three different proteins

Parvalbumin	V	K	K	A	F	A	I	I	D	Q	D	K	S	G	F	I	E	E	D	E	L	K	L	F	L	Q	N	F
Calmodulin	F	K	E	A	F	S	L	F	D	K	D	G	D	G	T	I	T	T	K	E	L	G	T	V	M	R	S	L
Troponin-C	L	A	D	C	F	R	I	F	D	K	N	A	D	G	F	I	D	I	E	E	L	G	E	I	L	R	A	T

Calcium-binding residues are orange, and residues that form the hydrophobic core of the motif are light green. The helix-loop-helix region shown underneath is colored as in Figure 2.13.

protein troponin-C as well as calmodulin, which regulates a variety of cellular functions by calcium binding. Subsequent structure determinations of these two proteins have shown that Kretsinger's prediction was correct. The structure of one of these, troponin-C, which was determined by Osnat Herzberg in the laboratory of Michael James in Edmonton, Canada, to 2.0 Å resolution, is shown in Figure 2.13. Functional aspects of calcium binding to calmodulin and the dramatic structural change of this molecule when it binds to target peptides are discussed in Chapter 6.

The hairpin β motif occurs frequently in protein structures

The simplest motif involving β strands, simpler than the α-helical calcium-binding motif, is two adjacent antiparallel strands joined by a loop. This motif, which is called either a hairpin or a β-β unit, occurs quite frequently; it is present in most antiparallel β structures both as an isolated ribbon and as part of more complex β sheets. There is a strong preference for β strands to be adjacent in β sheets when they are adjacent in the amino acid sequence and thus to form a **hairpin β** motif (β hairpin for short). The lengths of the loop regions between the β strands vary but are generally from two to five residues long (see Figure 2.8). There is no specific function associated with this motif.

Figure 2.14 shows examples of both cases, an isolated ribbon and a β sheet. The isolated ribbon is illustrated by the structure of bovine trypsin inhibitor (Figure 2.14a), a small, very stable polypeptide of 58 amino acids that inhibits the activity of the digestive protease trypsin. The structure has been determined to 1.0 Å resolution in the laboratory of Robert Huber in Munich, Germany, and the folding pathway of this protein is discussed in Chapter 6. Hairpin motifs as parts of a β sheet are exemplified by the structure of a snake venom, erabutoxin (Figure 2.14b), which binds to and inhibits

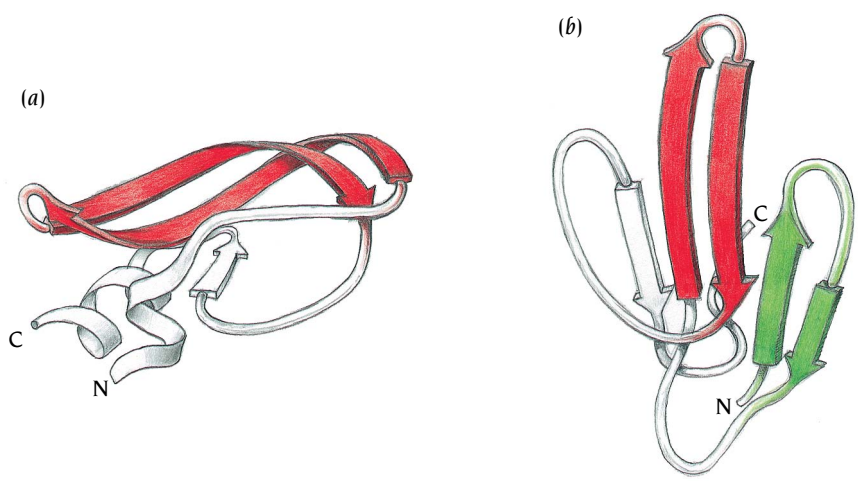


Figure 2.14 The hairpin motif is very frequent in β sheets and is built up from two adjacent β strands that are joined by a loop region. Two examples of such motifs are shown. (a) Schematic diagram of the structure of bovine trypsin inhibitor. The hairpin motif is colored red. (b) Schematic diagram of the structure of the snake venom erabutoxin. The two hairpin motifs within the β sheet are colored red and green. (Adapted from J. Richardson.)

Petitioner Merck, Ex. 1014, p. 26

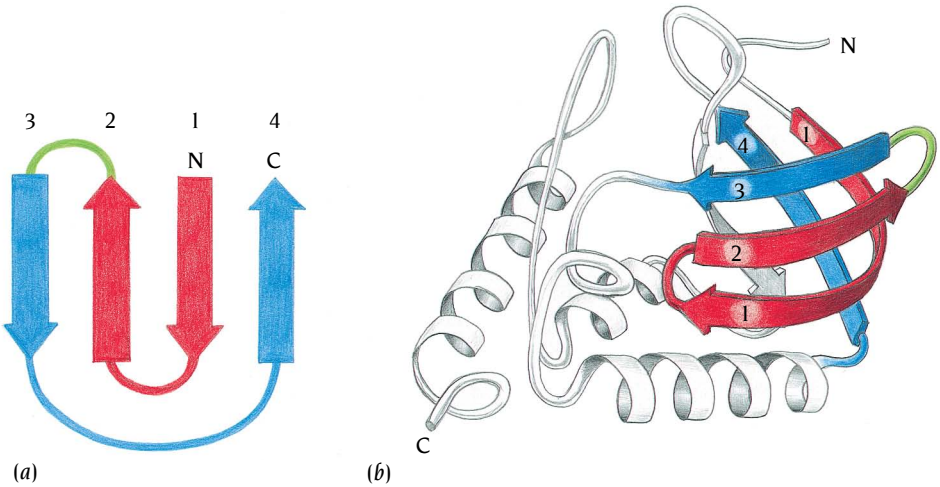


Figure 2.15 The Greek key motif is found in antiparallel β sheets when four adjacent β strands are arranged in the pattern shown as a topology diagram in (a). The motif occurs in many β sheets and is exemplified here by the enzyme *Staphylococcus* nuclease (b). The four β strands that form this motif are colored red and blue. The structure of this enzyme was determined to 1.5 Å resolution in the laboratory of Al Cotton at MIT. (Adapted from J. Richardson.)

the acetylcholine receptor in nerve cells. The structure has been determined to 1.4 Å resolution in the laboratory of Barbara Low at Columbia University. The core of this structure is a β sheet of five strands that contains two hairpin motifs and one additional β strand.

The Greek key motif is found in antiparallel β sheets

Four adjacent antiparallel β strands are frequently arranged in a pattern similar to the repeating unit of an ornamental pattern, or fret, used in ancient Greece, which is now called a Greek key. In proteins the motif is therefore called a **Greek key motif**, and Figure 2.15 shows an example of such a motif in the structure of *Staphylococcus* nuclease, an enzyme that degrades DNA. The Greek key motif is not associated with any specific function, but it occurs frequently in protein structures.

The Greek key motifs have been analyzed in detail, and it has been suggested that their frequent occurrence compared with other arrangements of four antiparallel β strands is based on an initial formation of one long antiparallel structure with loops in the middle of both β strands, as shown in Figure 2.16. By structural changes in the loop regions between β strands 1 and 2 and between β strands 3 and 4, the top part folds down so that β strand 2 associates with β strand 1. Beta strands 1 and 2 then form hydrogen bonds, and the Greek key motif is thus formed.

The β-α-β motif contains two parallel β strands

The hairpin motif is a simple and frequently used way to connect two antiparallel β strands, since the connected ends of the β strands are close together at the same edge of the β sheet. How are parallel β strands connected? If two adjacent strands are consecutive in the amino acid sequence, the two ends that must be joined are at opposite edges of the β sheet. The polypeptide chain must cross the β sheet from one edge to the other and connect the next β strand close to the point where the first β strand started. Such crossover connections are frequently made by α helices. The polypeptide chain must turn twice using loop regions, and the motif that is formed is thus a β strand followed by a loop, an α helix, another loop, and, finally, the second β strand.

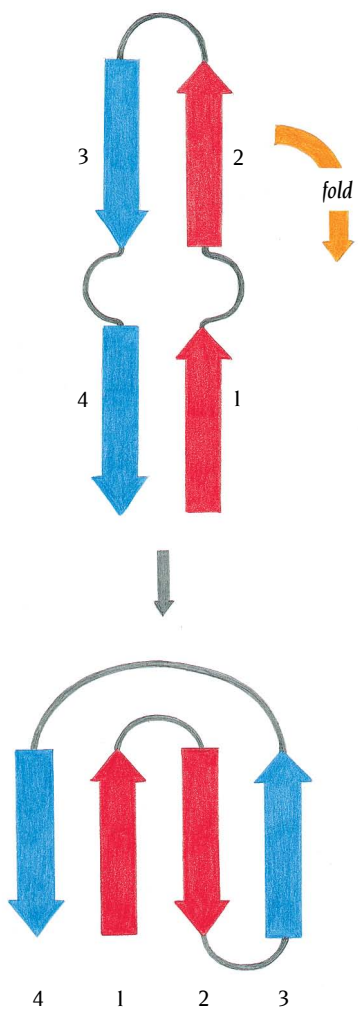


Figure 2.16 Suggested folding pathway from a hairpinlike structure to the Greek key motif. Beta strands 2 and 3 fold over such that strand 2 is aligned adjacent and antiparallel to strand 1. The topology diagram of the Greek key shown here is the same as in Figure 2.15a but rotated 180° in the plane of the page.

Petitioner Merck, Ex. 1014, p. 27

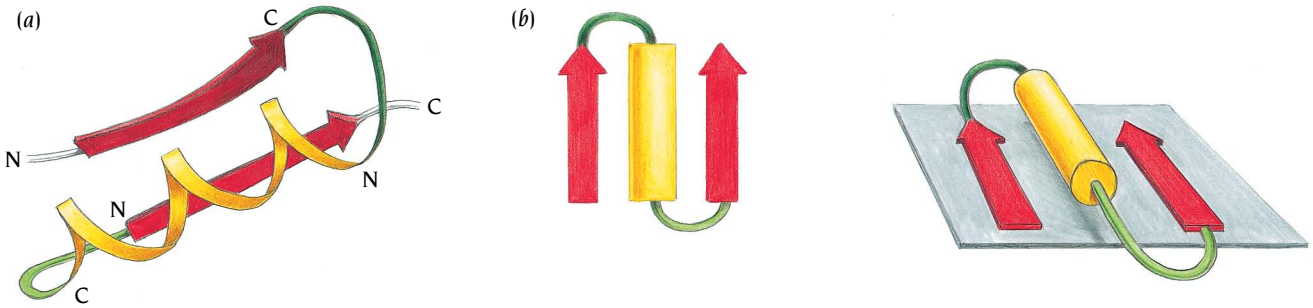


Figure 2.17 Two adjacent parallel β strands are usually connected by an α helix from the C-terminus of strand 1 to the N-terminus of strand 2. Most protein structures that contain parallel β sheets are built up from combinations of such β-α-β motifs. Beta strands are red, and α helices are yellow. Arrows represent β strands, and cylinders represent helices. (a) Schematic diagram of the path of the main chain. (b) Topological diagrams of the β-α-β motif.

This motif is called a **beta-alpha-beta motif** (Figure 2.17) and is found as part of almost every protein structure that has a parallel β sheet. For example, the molecule shown in Figure 2.10b, triosephosphate isomerase, is entirely built up by repeated combinations of this motif, where two successive motifs share one β strand. Alternatively, it can be regarded as being built up from four consecutive β-α-β-α motifs.

The α helix in the β-α-β motif connects the carboxy end of one β strand with the amino end of the next β strand (see Figure 2.17) and is usually oriented so that the helical axis is approximately parallel to the β strands. The α helix packs against the β strands and thus shields the hydrophobic residues of the β strands from the solvent. The β-α-β motif thus consists of two parallel β strands, an α helix, and two loop regions (except in a few cases where the connection between the two parallel β strands is not an α helix but a polypeptide chain of irregular structure). The loop regions can be of very different lengths, from one or two residues to over a hundred. The two loops have different functions. The loop (dark green in Figure 2.17) that connects the carboxy end of the β strand with the amino end of the α helix is often involved in forming the functional binding site, or active site, of these structures. These loop regions thus usually have conserved amino acid sequences in homologous proteins. In contrast, the other loop (light green in Figure 2.17) has not yet been found to contribute to an active site.

The β-α-β motif can be regarded as a loose helical turn from one β strand, around the connection, and into the next β strand. The motif can thus in principle have two different "hands" (Figure 2.18). Essentially every β-α-β motif in the known protein structures has been found to have the same hand as a right-handed α helix and therefore is called right-handed. No convincing explanation has been found for this regularity, even though it is the only general rule that describes how three secondary structure elements are arranged relative to each other. This handedness has important structural and functional consequences when several of these motifs are linked into a domain structure, as will be described in Chapter 4.

Protein molecules are organized in a structural hierarchy

The Danish biochemist Kai Linderstrøm-Lang coined the terms "primary," "secondary," and "tertiary" structure to emphasize the structural hierarchy in

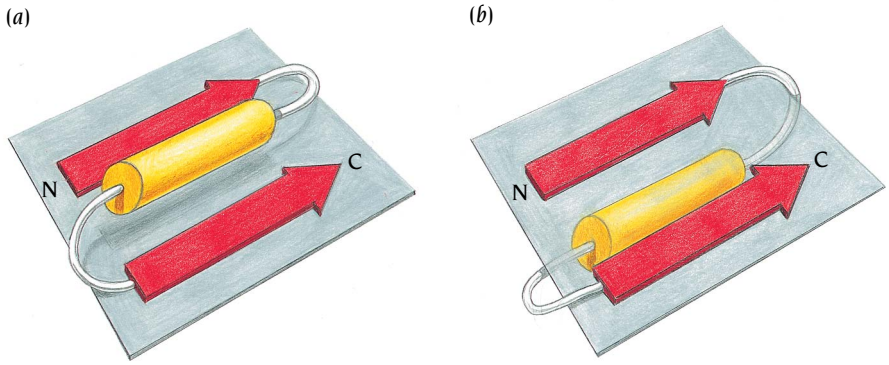


Figure 2.18 The β-α-β motif can in principle have two "hands." (a) This connection with the helix above the sheet is found in almost all proteins and is called right-handed because it has the same hand as a right-handed α helix. (b) The left-handed connection with the helix below the sheet.

Petitioner Merck, Ex. 1014, p. 28

proteins (see Figure 1.1). **Primary structure** is the amino acid sequence, or, in other words, the arrangement of amino acids along a linear polypeptide chain. Two different proteins that have significant similarities in their primary structures are said to be homologous to each other, and since their corresponding DNA sequences also are significantly similar, it is generally assumed that the two proteins are evolutionarily related, that they have evolved from a common ancestral gene.

Secondary structure occurs mainly as α helices and β strands. The formation of secondary structure in a local region of the polypeptide chain is to some extent determined by the primary structure. Certain amino acid sequences favor either α helices or β strands; others favor formation of loop regions. Secondary structure elements usually arrange themselves in simple motifs, as described earlier. Motifs are formed by packing side chains from adjacent α helices or β strands close to each other.

Several motifs usually combine to form compact globular structures, which are called **domains.** In this book we will use the term **tertiary structure** as a common term both for the way motifs are arranged into domain structures and for the way a single polypeptide chain folds into one or several domains. In all cases examined so far it has been found that if there is significant amino acid sequence homology in two domains in different proteins, these domains have similar tertiary structures.

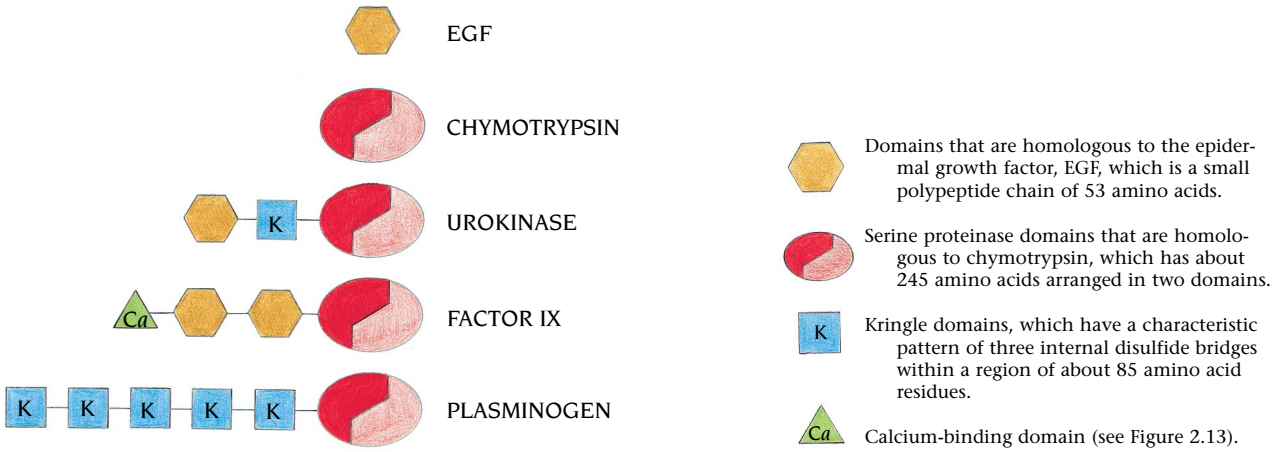
Protein molecules that have only one chain are called **monomeric** proteins. But a fairly large number of proteins have a **quaternary structure,** which consists of several identical polypeptide chains (subunits) that associate into a **multimeric** molecule in a specific way. These subunits can function either independently of each other or cooperatively so that the function of one subunit is dependent on the functional state of other subunits. Other protein molecules are assembled from several different subunits with different functions; for example, RNA polymerase from *E. coli* contains five different polypeptide chains.

Large polypeptide chains fold into several domains

The fundamental unit of tertiary structure is the domain. A domain is defined as a polypeptide chain or a part of a polypeptide chain that can fold independently into a stable tertiary structure. Domains are also units of function. Often, the different domains of a protein are associated with different functions. For example, in the lambda repressor protein, discussed in Chapter 8, one domain at the N-terminus of the polypeptide chain binds DNA, while a second domain at the C-terminus contains a site necessary for the dimerization of two polypeptide chains to form the dimeric repressor molecule.

Proteins may comprise a single domain or as many as several dozen domains (Figure 2.19). There is no fundamental structural distinction

Figure 2.19 Organization of polypeptide chains into domains. Small protein molecules like the epidermal growth factor, EGF, comprise only one domain. Others, like the serine proteinase chymotrypsin, are arranged in two domains that are required to form a functional unit (see Chapter 11). Many of the proteins that are involved in blood coagulation and fibrinolysis, such as urokinase, factor IX, and plasminogen, have long polypeptide chains that comprise different combinations of domains homologous to EGF and serine proteinases and, in addition, calcium-binding domains and Kringle domains.



Petitioner Merck, Ex. 1014, p. 29

between a domain and a subunit, there are many known examples where several biological functions that are carried out by separate polypeptide chains in one species are performed by domains of a single protein in another species. For example, synthesis of fatty acids requires catalysis of seven different chemical reactions. In plant chloroplasts these reactions are catalyzed by seven different proteins, whereas in mammals they are performed by one polypeptide chain arranged in seven domains with short linker regions between the domains. Such differences thus reflect the organization of the genome rather than the dictates of structure.

Domains are built from structural motifs

Domains are formed by different combinations of secondary structure elements and motifs. The α helices and β strands of the motifs are adjacent to each other in the three-dimensional structure and connected by loop regions. Sequentially adjacent motifs, or motifs that are formed from consecutive regions of the primary structure of a polypeptide chain, are usually close together in the three-dimensional structure (Figure 2.20). Thus to a first approximation a polypeptide chain can be considered as a sequential arrangement of these simple motifs. The number of such combinations found in proteins is limited, and some combinations seem to be structurally favored. Thus similar domain structures frequently occur in different proteins with different functions and with completely different amino acid sequences.

Simple motifs combine to form complex motifs

Figure 2.21 illustrates the 24 possible ways in which two adjacent β hairpin motifs, each consisting of two antiparallel β strands connected by a loop region, can be combined to make a more complex motif.

A survey of all known structures in 1991 showed that only those eight arrangements shown in Figure 2.21a occurred either as a complete β sheet or as a fragment of a β sheet with more than four strands. The number of times that these complex motifs occurred were 65, 29, 23, 11, 9, 3, 2, 1 for (i) to

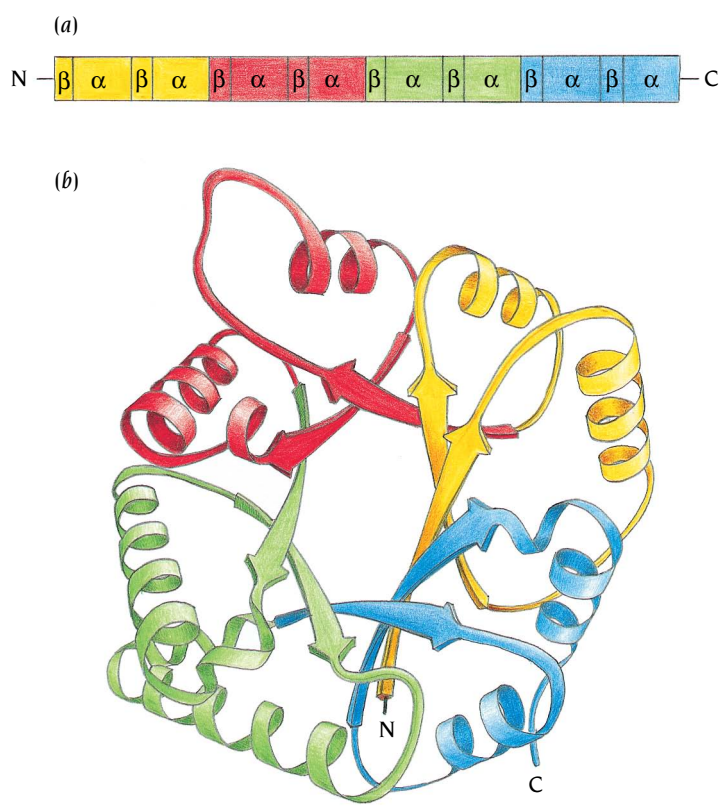


Figure 2.20 Motifs that are adjacent in the amino acid sequence are also usually adjacent in the three-dimensional structure. Triosephosphate isomerase is built up from four β-α-β-α motifs that are consecutive both in the amino acid sequence (a) and in the three-dimensional structure (b).

Petitioner Merck, Ex. 1014, p. 30

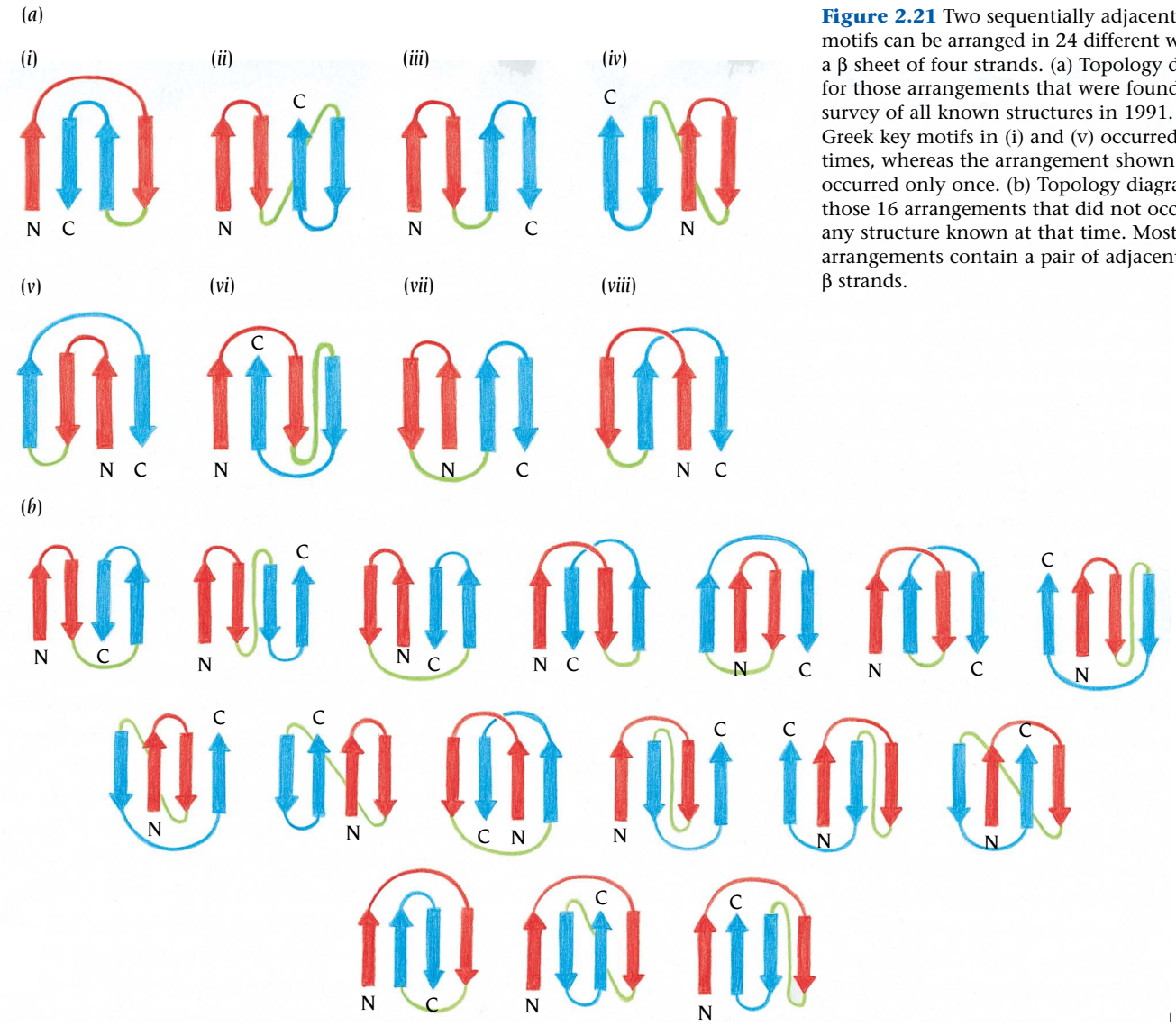


Figure 2.21 Two sequentially adjacent hairpin motifs can be arranged in 24 different ways into a β sheet of four strands. (a) Topology diagrams for those arrangements that were found in a survey of all known structures in 1991. The Greek key motifs in (i) and (v) occurred 74 times, whereas the arrangement shown in (viii) occurred only once. (b) Topology diagrams for those 16 arrangements that did not occur in any structure known at that time. Most of these arrangements contain a pair of adjacent parallel β strands.

(viii), respectively. The Greek key motifs shown in (i) and (v) thus occur more frequently than all the other possible motifs together. It is also apparent that two hairpin motifs strongly prefer to combine in such a way that all four strands become antiparallel, rather than being arranged with two adjacent parallel β strands. Of 143 β sheets with 4 strands, 138 were antiparallel and only 5 contained 2 adjacent parallel β strands.

The 16 possible motifs shown in Figure 2.21b never occur. Even if this database is limited compared with the universe of existing proteins, the survey clearly demonstrates that a few topological arrangements occur much more frequently than others, and that most possible complex motifs never occur or occur only in a few cases. In 1995 a preliminary survey of a much larger database of known structures yielded the same basic conclusions.

Protein structures can be divided into three main classes

On the basis of simple considerations of connected motifs, Michael Levitt and Cyrus Chothia of the MRC Laboratory of Molecular Biology derived a taxonomy of protein structures and have classified domain structures into three main groups: α domains, β domains, and α/β domains. In α structures the core is built up exclusively from α helices (see Figure 2.9); in β structures the core comprises antiparallel β sheets and are usually two β sheets packed

Petitioner Merck, Ex. 1014, p. 31

against each other (see Figure 2.11c). The α/β structures are made from combinations of β-α-β motifs that form a predominantly parallel β sheet surrounded by α helices (see Figures 2.10b and 2.11b).

Some proteins are built up from a combination of discrete α and β motifs and usually form one small antiparallel β sheet in one part of the domain packed against a number of α helices (see Figure 2.15). These structures can be considered to belong to a small fourth group called α + β. In addition to these groups, there are a number of small proteins that are rich in disulfide bonds or metal atoms and form a special group. The structures of these proteins seem to be strongly influenced by the presence of these metals or disulfides and often look like distorted versions of more regular proteins.

In this book the domains of the known protein structures are classified according to Levitt and Chothia's scheme. The three main classes—α, β, and α/β—will be examined in more detail in Chapters 3, 4, and 5. The group of Cyrus Chothia has constructed a database in which all available protein structures are arranged according to this hierarchy. Within each class the structures are arranged in superfamilies according to their tertiary structure and, within the superfamilies, in families according to function and sequence homology. This database is freely available on the World Wide Web.

Conclusion

The interiors of protein molecules contain mainly hydrophobic side chains. The main chain in the interior is arranged in secondary structures to neutralize its polar atoms through hydrogen bonds. There are two main types of secondary structure, α helices and β sheets. Beta sheets can have their strands parallel, antiparallel, or mixed.

Protein structures are built up by combinations of secondary structural elements, α helices, and β strands. These form the core regions—the interior of the molecule—and they are connected by loop regions at the surface. Schematic and simple topological diagrams where these secondary structure elements are highlighted are very useful and are frequently used. Alpha helices or β strands that are adjacent in the amino acid sequence are also usually adjacent in the three-dimensional structure. Certain combinations, called motifs, occur very frequently, including the helix-loop-helix motif and the hairpin motif. A DNA-binding helix-loop-helix motif and a calcium-binding helix-loop-helix motif, each with its own specific geometry and amino acid sequence requirements, are used in many different proteins.

The β-α-β motif, which consists of two parallel β strands joined by an α helix, occurs in almost all structures that have a parallel β sheet. Four antiparallel β strands that are arranged in a specific way comprise the Greek key motif, which is frequently found in structures with antiparallel β sheets.

Polypeptide chains are folded into one or several discrete units, domains, which are the fundamental functional and three-dimensional structural units. The cores of domains are built up from combinations of small motifs of secondary structure, such as α-loop-α, β-loop-β, or β-α-β motifs. Domains are classified into three main structural groups: α structures, where the core is built up exclusively from α helices; β structures, which comprise antiparallel β sheets; and α/β structures, where combinations of β-α-β motifs form a predominantly parallel β sheet surrounded by α helices.

Petitioner Merck, Ex. 1014, p. 32

Selected readings

General

Chothia, C. Principles that determine the structure of proteins. *Annu. Rev. Biochem.* 53: 537–572, 1984.

Doolittle, R.F. Proteins. *Sci. Am.* 253: 88–99, 1985.

Hardie, D.G., Coggins, J.R. *Multidomain Proteins: Structure and Evolution.* Amsterdam: Elsevier, 1986.

Janin, J., Chothia, C. Domains in proteins: definitions, location and structural principles. *Methods Enzymol.* 115: 420–430, 1985.

Klotz, I.M., et al. Quaternary structure of proteins. *Annu. Rev. Biochem.* 39: 25–62, 1970.

Lesk, A.M. Themes and contrasts in protein structures. *Trends Biochem. Sci.* 9: June V, 1984.

Levitt, M., Chothia, C. Structural patterns in globular proteins. *Nature* 261: 552–558, 1976.

Matthews, B.W., Bernhard, S.A. Structure and symmetry of oligomeric enzymes. *Annu. Rev. Biophys. Bioeng.* 2: 257–317, 1973.

Richardson, J.S. Describing patterns of protein tertiary structure. *Methods Enzymol.* 115: 349–358, 1985.

Richardson, J.S. Schematic drawings of protein structures. *Methods Enzymol.* 115: 359–380, 1985.

Richardson, J.S. The anatomy and taxonomy of protein structure. *Adv. Prot. Chem.* 34: 167–339, 1981.

Rossmann, M.G., Argos, P. Protein folding. *Annu. Rev. Biochem.* 50: 497–532, 1981.

Schulz, G.E. Protein differentiation: emergence of novel proteins during evolution. *Angew. Chem.,* int. ed. 20: 143–151, 1981.

Schulz, G.E. Structural rules for globular proteins. *Angew. Chem.,* int. ed. 16: 23–33, 1977.

Strynadka, N.C.J., James, M.N.G. Crystal structures of the helix-loop-helix calcium-binding proteins. *Annu. Rev. Biochem.* 58: 951–998, 1989.

Specific structures

Adams, M.J., et al. Structure of lactate dehydrogenase at 2.8 Å resolution. *Nature* 227: 1098–1103, 1970.

Baba, Y.S., et al. Three-dimensional structure of calmodulin. *Nature* 315: 37–40, 1985.

Banner, B.W., et al. Structure of chicken muscle triose phosphate isomerase determined crystallographically at 2.5 Å resolution using amino acid sequence data. *Nature* 255: 609–614, 1975.

Bourne, P.E., et al. Erabutoxin b. Initial protein refinement and sequence analysis at 0.140 nm resolution. *Eur. J. Biochem.* 153: 521–527, 1985.

Burnett, R.M., et al. The structure of oxidized form of clostridial flavodoxin at 1.9 Å resolution. *J. Biol. Chem.* 249: 4383–4392, 1974.

Chothia, C. Structural invariants in protein folding. *Nature* 254: 304–308, 1975.

Chothia, C., Levitt, M., Richardson, D. Structure of proteins: packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci. USA* 74: 4130–4134, 1977.

Chou, P.Y., Fasman, G.D. β-turns in proteins. *J. Mol. Biol.* 115: 135–175, 1977.

Colman, P., et al. X-ray crystal structure analysis of plastocyanin at 2.7 Å resolution. *Nature* 272: 319–324, 1978.

Crawford, J.L., Lipscomb, W.N., Schellmann, C.G. The reverse turn as a polypeptide conformation in globular proteins. *Proc. Natl. Acad. Sci. USA* 70: 538–542, 1973.

Efimov, A.V. Stereochemistry of α-helices and β-sheet packing in compact globule. *J. Mol. Biol.* 134: 23–40, 1979.

Eklund, H., et al. Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution. *J. Mol. Biol.* 102: 27–59, 1976.

Gouaux, J.E., Lipscomb, W.N. Crystal structures of phosphonoacetamide ligated T and phosphonoacetamide and malonate ligated R states of aspartate carbamoyltransferase at 2.8 Å resolution and neutral pH. *Biochemistry* 29: 389–402, 1990.

Herzberg, O., James, M.N.G. Structure of the calcium regulatory muscle protein troponin-C at 2.8 Å resolution. *Nature* 313: 653–659, 1985.

Hol, W.G.J., van Duijnen, P.T., Berendsen, H.J.C. The α-helix dipole and the properties of proteins. *Nature* 273: 443–446, 1978.

Holmgren, A., et al. Three-dimensional structure of *E. coli* thioredoxin-S_2 to 2.8 Å resolution. *Proc. Natl. Acad. Sci. USA* 72: 2305–2309, 1975.

Jones, A., Thirup, S. Using known substructures in protein model building and crystallography. *EMBO J.* 5: 819–822, 1986.

Kendrew, J.C. The three-dimensional structure of a protein molecule. *Sci. Am.* 205: 96–110, 1961.

Kendrew, J.C., et al. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181: 662–666, 1958.

Kendrew, J.C., et al. Structure of myoglobin. *Nature* 185: 422–427, 1960.

Koch, I., Kaden, F., Selbig, J. Analysis of protein sheet topologies by graph theoretical methods. *Prot: Struc. Func. Genet.* 12: 314–323, 1992.

Kretsinger, R.H. Structure and evolution of calcium-modulated proteins. *CRC Crit. Rev. Biochem.* 8: 119–174, 1980.

Lesk, A.M., Hardman, K.D. Computer-generated pictures of proteins. *Methods Enzymol.* 115: 381–390, 1985.

Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 17: 4277–4285, 1978.

Matthews, B.W., Rossmann, M.G. Comparison of protein structures. *Methods Enzymol.* 115: 397–420, 1985.

Petitioner Merck, Ex. 1014, p. 33

Milner-White, E.J., Poet, R. Loops, bulges, turns and hairpins in proteins. *Trends Biochem. Sci.* 12: 189–192, 1987.

Moews, P.C., Kretsinger, R.H. Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. *J. Mol. Biol.* 91: 201–228, 1975.

Park, C.H., Tulinsky, A. Three-dimensional structure of the Kringle sequence: structure of prothrombin fragment 1. *Biochemistry* 25: 3977–3982, 1986.

Pauling, L., Corey, R.B. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc. Natl. Acad. Sci. USA* 37: 729–740, 1951.

Pauling, L., Corey, R.B., Branson, H.R. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* 37: 205–211, 1951.

Perutz, M.F. Electrostatic effects in proteins. *Science* 201: 1187–1191, 1978.

Perutz, M.F. New x-ray evidence on the configuration of polypeptide chains. Polypeptide chains in poly-g-benzyl-t-glutamate, keratin and haemoglobin. *Nature* 167: 1053–1054, 1951.

Perutz, M.F., et al. Structure of haemoglobin. A three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by x-ray analysis. *Natur*e 185: 416–422, 1960.

Rao, S.T., Rossmann, M.G. Comparison of super-secondary structures in proteins. *J. Mol. Biol.* 76: 241–256, 1973.

Remington, S.J., Matthews, B.W. A systematic approach to the comparison of protein structures. *J. Mol. Biol.* 140: 77–99, 1980.

Richards, F.M. Calculation of molecular volumes and areas for structures of known geometry. *Methods Enzymol.* 115: 440–464, 1985.

Rose, G.D. Automatic recognition of domains in globular proteins. *Methods Enzymol.* 115: 430–440, 1985.

Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* 272: 586–590, 1978.

Rose, G.D., Roy, S. Hydrophobic basis of packing in globular proteins. *Proc. Natl. Acad. Sci. USA* 77: 4643–4647, 1980.

Rose, G.D., Young, W.B., Gierasch, L.M. Interior turns in globular proteins. *Nature* 304: 654–657, 1983.

Sibanda, B.L., Thornton, J.M. β-hairpin families in globular proteins. *Nature* 316: 170–174, 1985.

Tucker, P.W., Hazen, E.E., Cotton, F.A. Staphylococcal nuclease reviewed: a prototypic study in contemporary enzymology. III. Correlation of the three-dimensional structure with the mechanisms of enzymatic action. *Mol. Cell. Biochem.* 23: 67–86, 1979.

Venkatachalam, C.M. Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers* 6: 1425–1436, 1968.

Wiegand, G., et al. Crystal structure analysis and molecular model of a complex of citrate synthase with oxaloacetate and S-acetonyl-coenzyme A. *J. Mol. Biol.* 174: 205–219, 1984.

Wlodawer, A., Deisenhofer, J., Huber, R. Comparison of two highly refined structures of bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* 193: 145–156, 1987.

Wright, C.S., Alden, R., Kraut, J. Structure of subtilisin BPN′ at 2.5 Å resolution. *Nature* 221: 235–242, 1969.

Petitioner Merck, Ex. 1014, p. 34

Alpha-Domain Structures

3

The first globular protein structure that was determined, myoglobin, belongs to the class of alpha- (α-) domain structures. The structure illustrated in Figure 2.9 is called the globin fold and is a representative example of one class of α domains in proteins; short α helices, the building blocks, are connected by loop regions and packed together to produce a hydrophobic core. Packing interactions within the core hold the helices together in a stable globular structure, while the hydrophilic residues on the surface make the protein soluble in water. In this chapter we will describe some of the different α-domain structures in soluble proteins.

Alpha helices are sufficiently versatile to produce many very different classes of structures. In membrane-bound proteins, the regions inside the membranes are frequently α helices whose surfaces are covered by hydrophobic side chains suitable for the hydrophobic environment inside the membranes. Membrane-bound proteins are described in Chapter 12. Alpha helices are also frequently used to produce structural and motile proteins with various different properties and functions. These can be typical fibrous proteins such as keratin, which is present in skin, hair, and feathers, or parts of the cellular machinery such as fibrinogen or the muscle proteins myosin and dystrophin. These α-helical proteins will be discussed in Chapter 14.

Coiled-coil α helices contain a repetitive heptad amino acid sequence pattern

Despite its frequent occurrence in proteins an isolated α helix is only marginally stable in solution. Alpha helices are stabilized in proteins by being packed together through hydrophobic side chains. The simplest way to achieve such stabilization is to pack two α helices together. As early as 1953 Francis Crick showed that the side-chain interactions are maximized if the two α helices are not straight rods but are wound around each other in a supercoil, a so-called **coiled-coil** arrangement (Figure 3.1). Coiled-coils are the basis for some of the fibrous proteins we shall discuss in Chapter 14. Coiled-coils in fibers can extend over many hundreds of amino acid residues

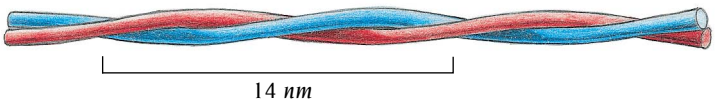


Figure 3.1 Schematic diagram of the coiled-coil structure. Two α helices are intertwined and gradually coil around each other.

Petitioner Merck, Ex. 1014, p. 35

	a	b	c	d	e	f	g	
NH_2 -	Met	- Lys	- Gln	- **Leu**	- Glu	- Asp	- Lys	-
	Val	- Glu	- Glu	- **Leu**	- Leu	- Ser	- Lys	-
	Asn	- Tyr	- His	- **Leu**	- Glu	- Asn	- Glu	-
	Val	- Ala	- Arg	- **Leu**	- Lys	- Lys	- Leu	- COOH

(a)

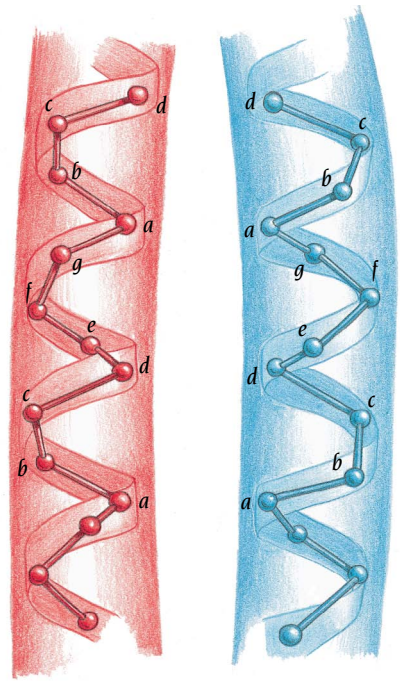


Figure 3.2 Repetitive pattern of amino acids in a coiled-coil α helix. (a) The amino acid sequence of the transcription factor GCN4 showing a heptad repeat of leucine residues. Within each heptad the amino acids are labeled a–g. (b) Schematic diagram of one heptad repeat in a coiled-coil structure showing the backbone of the polypeptide chain. The α helices in the coiled-coil are slightly distorted so that the helical repeat is 3.5 residues rather than 3.6, as in a regular helix. There is therefore an integral repeat of seven residues along the helix.

to produce long, flexible dimers that contribute to the strength and flexibility of the fibers. Much shorter coiled-coils are used in some transcription factors to promote or prevent formation of homo- and heterodimers, as we shall discuss in Chapter 10.

Crick showed that a left-handed supercoil of two right-handed α helices reduces the number of residues per turn in each helix from 3.6 to 3.5 so that the pattern of side-chain interactions between the helices repeats every seven residues, that is, after two turns. This is reflected in the amino acid sequences of polypeptide chains that form α-helical coiled-coils. Such sequences are repetitive with a period of seven residues, the **heptad repeat.** The amino acid residues within one such heptad repeat are usually labeled a–g (Figure 3.2a), and one of these, the d-residue, is hydrophobic, usually a leucine or an isoleucine. When two α helices form a coiled-coil structure the side chains of these d-residues pack against each other every second turn of the α helices (Figure 3.2b). The hydrophobic region between the α helices is completed by the a-residues, which are frequently hydrophobic and also pack against each other (Figure 3.3). Residues "e" and "g," which border the hydrophobic core (see Figure 3.2b), frequently are charged residues. The side chains of these residues provide ionic interactions (salt bridges) between the α helices that define the relative chain alignment and orientation (Figure 3.4).

The repetitive heptad amino acid sequence pattern required for a coiled-coil structure can be identified in computer searches of amino acid sequence databases. Heptad repeats provide strong indications of α-helical coiled-coil structures, and they have been found in a number of different proteins with very diverse functions. Fibrinogen, which plays an essential role in blood coagulation; some RNA- and DNA-binding proteins; the class of cell-surface recognition proteins called collectins; both spectrin and dystrophin, which link actin molecules; and the muscle protein myosin all contain heptad repeats and therefore coiled-coil α helices. An illustrative example is provided by GCN4, a DNA-binding protein. GCN4 contains one region of α helix, the leucine zipper region, and its dimerization is accomplished by the formation of an α-helical coiled-coil with the leucine zipper regions of two subunits. The structure and DNA-binding function of this protein are described in Chapter 10.

Detailed structure determinations of GCN4 and other coiled-coil proteins have shown that the α helices pack against each other according to the "knobs in holes" model first suggested by Francis Crick (Figure 3.5). Each side chain in the hydrophobic region of one of the α helices can contact four side chains from the second α helix. The side chain of a residue in position "d"

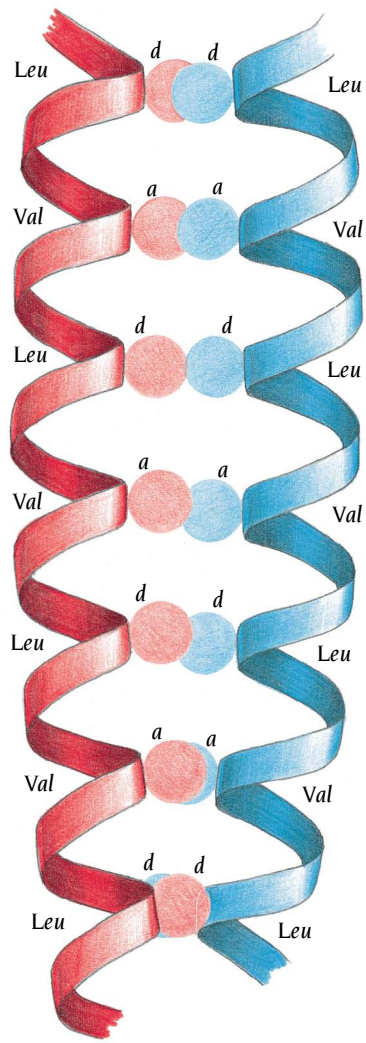


Figure 3.3 Schematic diagram showing the packing of hydrophobic side chains between the two α helices in a coiled-coil structure. Every seventh residue in both α helices is a leucine, labeled "d." Due to the heptad repeat, the d-residues pack against each other along the coiled-coil. Residues labeled "a" are also usually hydrophobic and participate in forming the hydrophobic core along the coiled-coil.

Petitioner Merck, Ex. 1014, p. 36

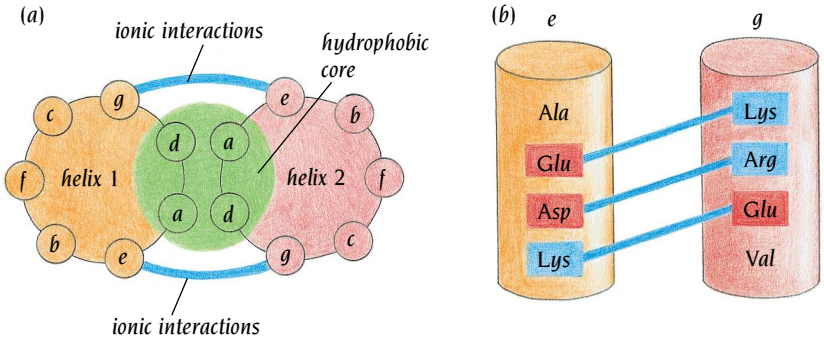


Figure 3.4 Salt bridges can stabilize coiled-coil structures and are sometimes important for the formation of heterodimeric coiled-coil structures. The residues labeled "e" and "g" in the heptad sequence are close to the hydrophobic core and can form salt bridges between the two α helices of a coiled-coil structure, the e-residue in one helix with the g-residue in the second and vice versa. (a) Schematic view from the top of a heptad repeat. (b) Schematic view from the side of a coiled-coil structure.

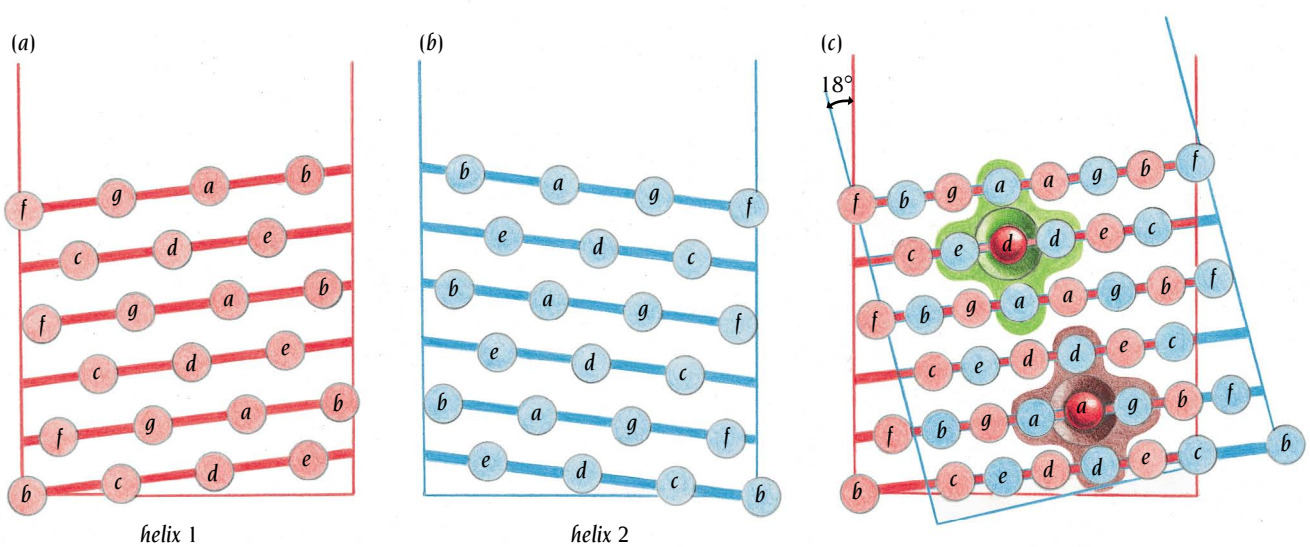
in one helix is directed into a hole at the surface of the second helix surrounded by one d-residue, two a-residues, and one e-residue, with numbers n, $n - 3$, $n + 4$, and $n + 1$, respectively. The two helices are aligned in such a way that the two d-residues, frequently leucines or isoleucines, face each other (see Figure 3.3).

The four-helix bundle is a common domain structure in α proteins

Two α helices packed together into a coiled-coil are building blocks within a domain or a fiber but are not sufficient to form a complete domain. The simplest and most frequent α-helical domain consists of four α helices arranged in a bundle with the helical axes almost parallel to each other. A schematic representation of the structure of the **four-helix bundle** is shown in Figure 3.6a. The side chains of each helix in the four-helix bundle are arranged so that hydrophobic side chains are buried between the helices and hydrophilic side chains are on the outer surface of the bundle (Figure 3.6b). This arrangement creates a hydrophobic core in the middle of the bundle along its length, where the side chains are so closely packed that water is excluded.

The four-helix bundle occurs in several widely different proteins, such as myohemerythrin (an oxygen-transport protein in marine worms that does not contain heme iron), cytochrome c′ and cytochrome b_{562} (heme-containing electron carriers) (Figure 3.7a), ferritin (a storage molecule for iron atoms in eucaryotic cells), and the coat protein of tobacco mosaic virus. In these examples, sequentially adjacent α helices are always antiparallel. However, four-helix bundles can also be formed with different topological arrangements of the α helices. In human growth hormone (Figure 3.7b), a four-helix bundle is formed from two pairs of parallel α helices that are joined in an

Figure 3.5 Schematic diagram of packing side chains in the hydrophobic core of coiled-coil structures according to the "knobs in holes" model. The positions of the side chains along the surface of the cylindrical α helix is projected onto a plane parallel with the helical axis for both α helices of the coiled-coil. (a) Projected positions of side chains in helix 1. (b) Projected positions of side chains in helix 2. (c) Superposition of (a) and (b) using the relative orientation of the helices in the coiled-coil structure. The side-chain positions of the first helix, the "knobs," superimpose between the side-chain positions in the second helix, the "holes." The green shading outlines a d-residue (leucine) from helix 1 surrounded by four side chains from helix 2, and the brown shading outlines an a-residue (usually hydrophobic) from helix 1 surrounded by four side chains from helix 2.



Petitioner Merck, Ex. 1014, p. 37

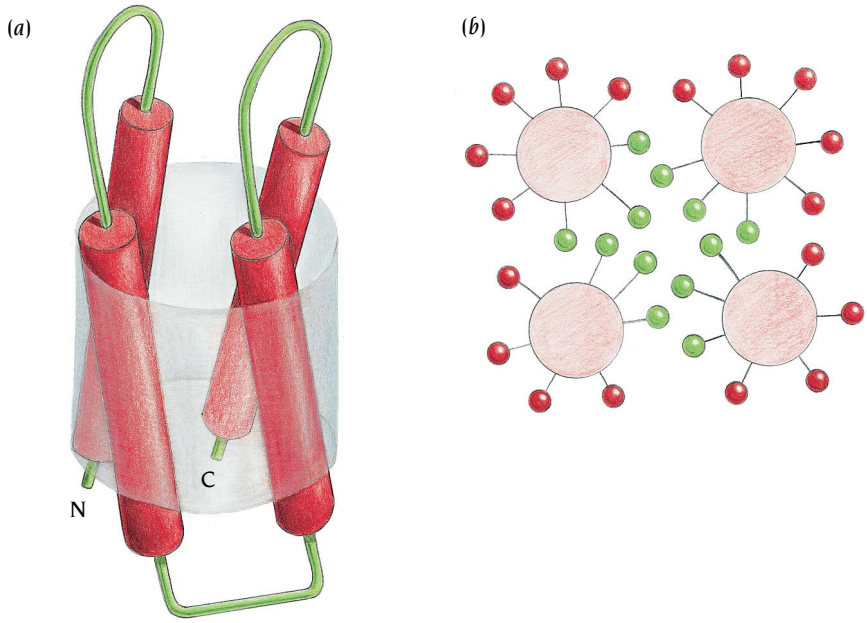


Figure 3.6 Four-helix bundles frequently occur as domains in α proteins. The arrangement of the α helices is such that adjacent helices in the amino acid sequence are also adjacent in the three-dimensional structure. Some side chains from all four helices are buried in the middle of the bundle, where they form a hydrophobic core. (a) Schematic representation of the path of the polypeptide chain in a four-helix-bundle domain. Red cylinders are α helices. (b) Schematic view of a projection down the bundle axis. Large circles represent the main chain of the α helices; small circles are side chains. Green circles are the buried hydrophobic side chains; red circles are side chains that are exposed on the surface of the bundle, which are mainly hydrophilic. [(a) Adapted from P.C. Weber and F.R. Salemme, *Nature* 287: 82–84, 1980.]

antiparallel fashion. The interaction of this hormone with its receptor is described in Chapter 13.

In most four-helix bundle structures, including those shown in Figure 3.7, the α helices are packed against each other according to the "ridges in grooves" model discussed later in this chapter. However, there are also examples where coiled-coil dimers packed by the "knobs in holes" model participate in four-helix bundle structures. A particularly simple illustrative example is the Rop protein, a small RNA-binding protein that is encoded by certain plasmids and is involved in plasmid replication. The monomeric subunit of Rop is a polypeptide chain of 63 amino acids built up from two

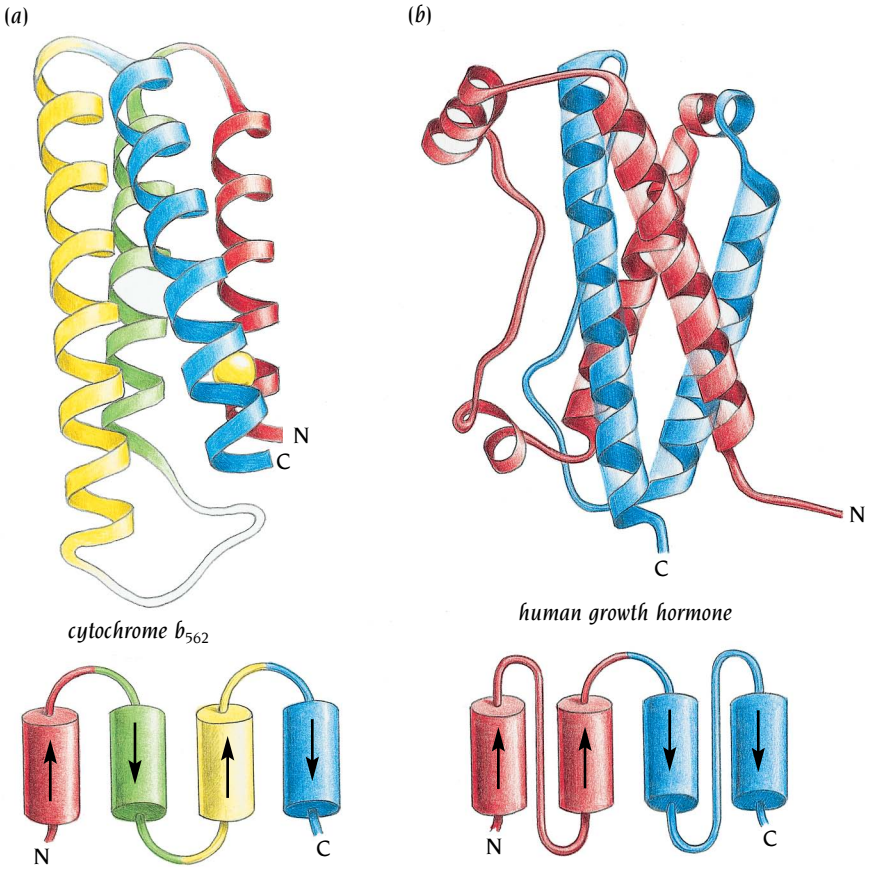


Figure 3.7 The polypeptide chains of cytochrome b_{562} and human growth hormone both form four-helix-bundle structures. In cytochrome b_{562} (a) adjacent helices are antiparallel, whereas the human growth hormone (b) has two pairs of parallel α helices joined in an antiparallel fashion.

Petitioner Merck, Ex. 1014, p. 38

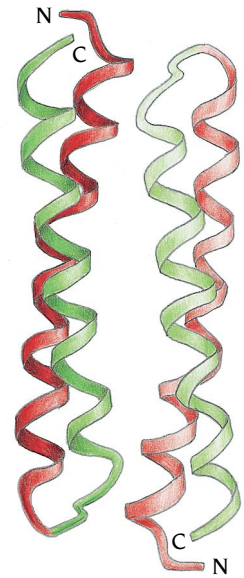


Figure 3.8 Schematic diagram of the dimeric Rop molecule. Each subunit comprises two α helices arranged in a coiled-coil structure with side chains packed into the hydrophobic core according to the "knobs in holes" model. The two subunits are arranged in such a way that a bundle of four α helices is formed.

antiparallel α helices joined by a short loop of three amino acids. The structure of Rop was determined by David Banner at EMBL, Heidelberg, Germany.

The two α helices of the Rop subunit are arranged as an antiparallel coiled-coil in which the hydrophobic side chains are packed against each other according to the "knobs in holes" model. Two such subunits, each with the same structure, form the dimeric Rop molecule in which the subunits are arranged as a bundle of four α helices with their long axes aligned (Figure 3.8). The two dimers pack against each other according to the "ridges in grooves" model. The helix-loop-helix (HLH) family of transcription factors, discussed in Chapter 10, is another example of a four-helix bundle structure involving coiled-coil helices.

Alpha-helical domains are sometimes large and complex

The structures of several enzymes are known in which a long polypeptide chain of 300–400 amino acids is arranged in more than 20 α helices packed together in a complex pattern to form a globular domain. One such enzyme is a bacterial muramidase that is involved in the metabolism of peptidoglycans, which form part of the bacterial cell wall. The structure of this enzyme was determined by Bauke Dijkstra and colleagues in Groningen, Netherlands, as a basis for the design of specific inhibitors to the enzyme, which might lead eventually to novel types of antibacterial drugs.

The polypeptide chain of this monomeric enzyme has 618 amino acids, of which the N-terminal 450 residues form one α-helical domain. This domain is built up from 27 α helices arranged in a two-layered ring with a right-handed superhelical twist (Figure 3.9). The ring has a large central hole, like in a doughnut, with a diameter of about 30 Å. The remaining residues form the catalytic domain that lies on top of the ring. The function of the

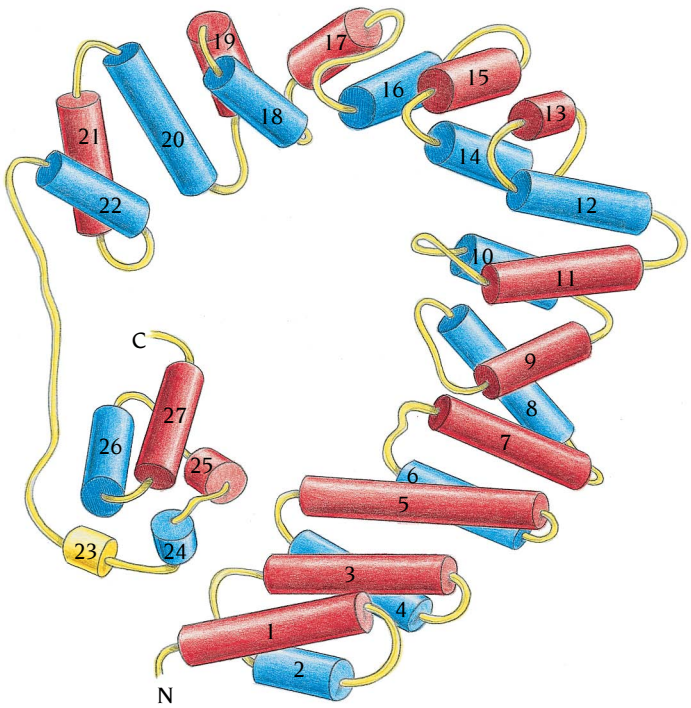


Figure 3.9 Schematic diagram of the structure of one domain of a bacterial muramidase, comprising 450 amino acid residues. The structure is built up from 27 α helices arranged in a two-layered ring. The ring has a large central hole, like a doughnut, with a diameter of about 30 Å.

Petitioner Merck, Ex. 1014, p. 39

doughnut-shaped domain is not known, but its shape may be required for the specificity of the catalytic reaction *in vivo*.

The globin fold is present in myoglobin and hemoglobin

One of the most important α structures is the **globin fold.** This fold has been found in a large group of related proteins, including myoglobin, hemoglobins, and the light-capturing assemblies in algae, the phycocyanins. The functional and evolutionary aspects of these structures will not be discussed in this book; instead, we will examine some features that are of general structural interest.

The pairwise arrangements of the sequential α helices in the globin fold are quite different from the antiparallel organization found in the four-helix-bundle α structures. The globin structure is a bundle of eight α helices, usually labeled A–H, connected by rather short loop regions and arranged so that the helices form a pocket for the active site, which in myoglobin and the hemoglobins binds a heme group (Figure 3.10). The lengths of the α helices vary considerably, from 7 residues in the shortest helix (C) to 28 in the longest helix (H) in myoglobin. In the globin fold the α helices wrap around the core in different directions so that sequentially adjacent α helices are usually not adjacent to each other in the structure. The only exceptions are the last two α helices (G and H), which form an antiparallel pair with extensive packing interactions between them. All other packing interactions are formed between pairs of α helices that are not sequentially adjacent. Because the globin fold is not built up from an assembly of smaller motifs, it is quite difficult to visualize conceptually in spite of its relatively small size and simplicity.

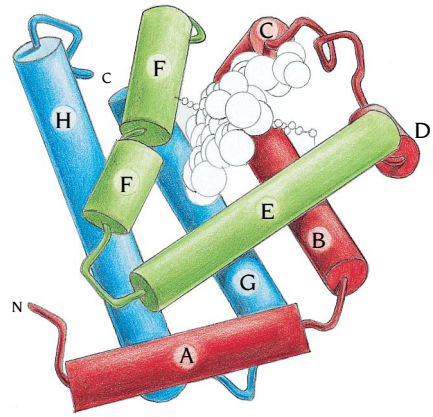


Figure 3.10 Schematic diagram of the globin domain. The eight α helices are labeled A–H. A–D are red, E and F green, and G and H blue. The heme group is shown in white. (Adapted from originals provided by A. Lesk.)

Geometric considerations determine α-helix packing

When we compare the arrangements of the α helices in coiled-coil structures (see Figure 3.1), in the four-helix-bundle structure (see Figure 3.8), and in the globin fold (see Figure 3.10), it is obvious that the geometry of α-helix packing is quite different. We described earlier the way that the side chains of coiled-coil α helices pack according to the "knobs in holes" model. In contrast, other α-helical structures pack their α helices according to a "ridges in grooves" model. In the four-helix bundle the α helices pack almost parallel, or antiparallel, to each other, with an angle of about 20° between the helical axes. In the globin fold the angles between the helical axes are usually larger, in most cases around 50°. These are the two main ways that α helices pack against each other in the "ridges in grooves" model, a packing motif dictated by the geometry of the surfaces of α helices.

Ridges of one α helix fit into grooves of an adjacent helix

Since the side chains of an α helix are arranged in a helical row along the surface of the helix, they form ridges separated by shallow furrows, or grooves, on the surface. Alpha helices pack with the ridges on one helix packing into the grooves of the other and vice versa. The ridges and grooves are formed by amino acids that are usually three or four residues apart. This is illustrated in Figure 3.11, which shows slices through the surface of a polyalanine α helix on which the directions of the ridges are marked. In contrast to the ridges and grooves of the DNA double helix described in Chapter 7, which are formed by the sugar-phosphate main-chain atoms, those of an α helix are formed by the amino acid side chains. The detailed geometry of the ridges and grooves of an α helix is thus dependent not only on the geometry of the helix but also on the actual amino acid sequence.

The most common way of packing α helices is by fitting the ridges formed by a row of residues separated in sequence by four in one helix into the same type of grooves in the other helix. In this case the ridges and

Petitioner Merck, Ex. 1014, p. 40

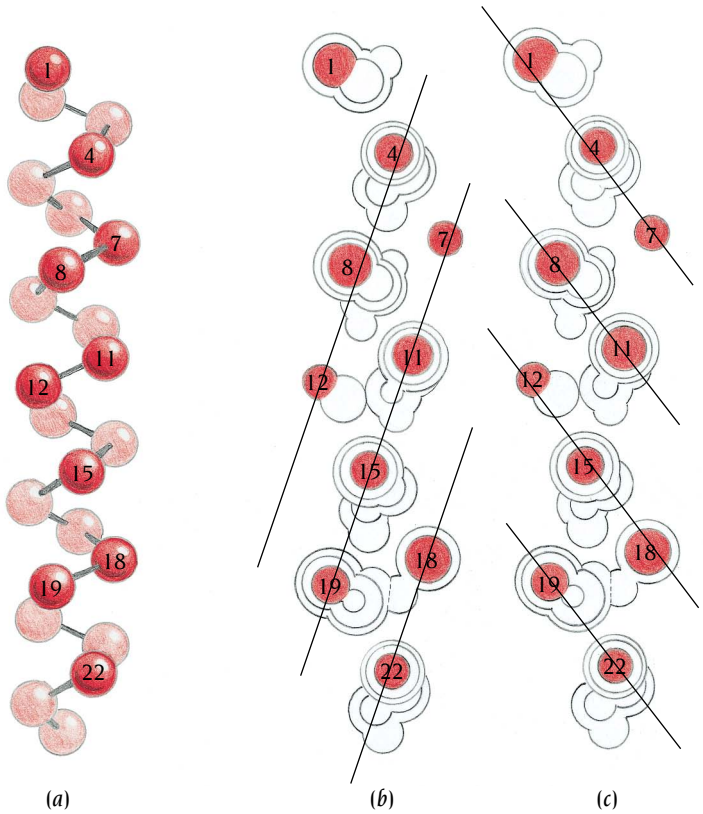


Figure 3.11 The side chains on the surface of an α helix form ridges separated by grooves, as schematically illustrated here. (a) An α helix with each residue represented by the first atom in the side chain, C_β. (b) The surface relief of a polyalanine α helix in the orientation shown in (a). Sections are cut through a space-filling model and superimposed. The residue numbers are placed on the side-chain atom. The ridges caused by the side chains separated by four residues are shown as lines. (c) The same as (b), but here the ridges are caused by side chains separated by three residues.

grooves form an angle of about 25° to the helical axis. In order to pack the two helices shown in Figure 3.12a (red and blue) against each other, one of these (the blue in Figure 3.12a) must be turned around 180° out of the plane of the paper and placed on top of the other (red). In the interface between the two α helices the directions of the ridges and grooves are then on opposite sides of the vertical axis, as illustrated in Figure 3.12a. The α helices must thus be inclined by an angle of about 50° (25° + 25°) in order for the ridges of one helix to fit into the grooves of the other and vice versa. This is the type of packing of several of the helix-helix interactions in the globin fold, and in many other helical structures.

In the second frequently occurring packing mode the ridges formed by amino acids three residues apart fit into the grooves of amino acids four residues apart and vice versa. The direction of the first type of ridge forms an angle of about 45° to the helical axis, whereas the other type makes an angle of about 25° to the axis in the opposite direction (Figure 3.12b). In the interface, however, after one helix has been rotated 180°, these directions are on the same side of the helical axis. Thus an inclination of about 20° (45°–25°) between the two α helices will fit these ridges and grooves into each other. Some four-helix-bundle structures (see Figure 3.8b) have this mode of packing.

These two rules for fitting ridges into grooves are quite general: they apply to most packing interactions between α helices, and they explain the geometrical arrangements of adjacent α helices observed in many protein structures.

The globin fold has been preserved during evolution

The three-dimensional structures of globin domains from many diverse sources, including mammals, insects, and plant root nodules, have been determined independently of each other. All these domains have amino acid

Petitioner Merck, Ex. 1014, p. 41

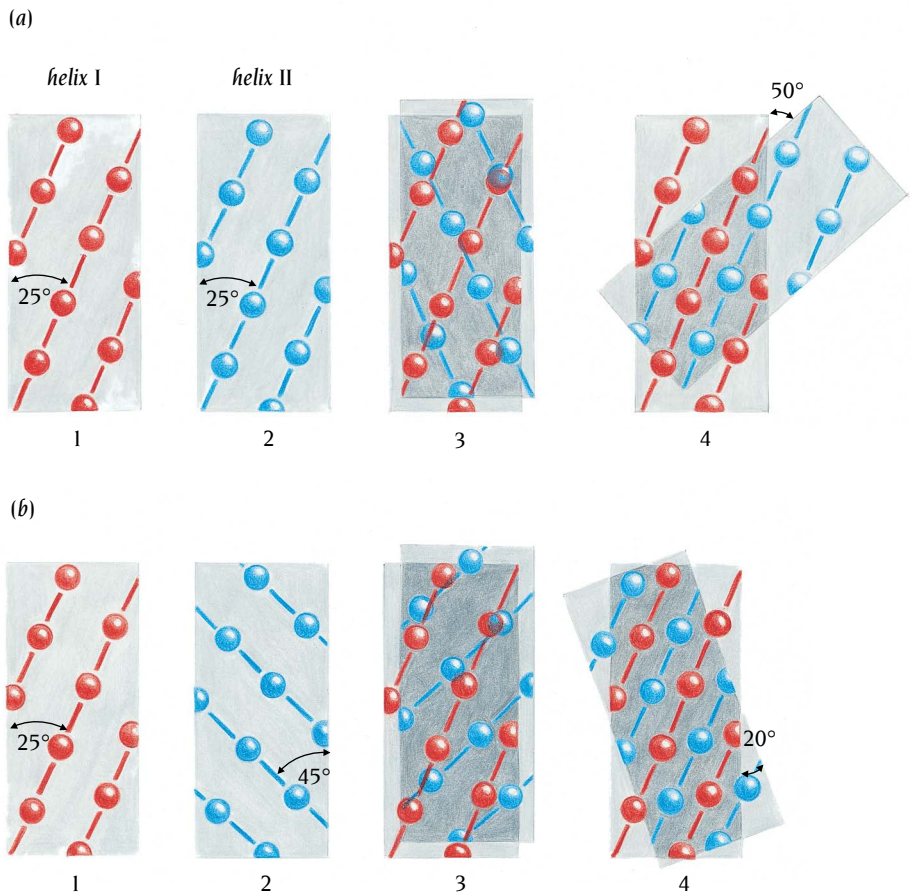


Figure 3.12 By fitting the ridges of side chains from one helix into the grooves between side chains of the other helix and vice versa, α helices pack against each other. (a) Two α helices, I and II, with ridges from side chains separated by four residues marked in red and blue, respectively. Panels 1 and 2 are the same view of the two α helices. In panel 3 the blue α helix is turned over through 180° in order to form an interface with the red α helix. In panel 4 the orientation of the helices has been rotated 50° in order to pack the ridges of one α helix into the grooves of the other. (b) In the red α helix the ridges are formed by side chains separated by four residues and in the blue α helix by three residues. The α helices are rotated 20° in order to pack ridges into grooves, in a direction opposite that in (a). (Adapted from C. Chothia et al., *Proc. Natl. Acad. Sci. USA* 74: 4130–4134, 1977.)

sequence homologies that range from 99% to 16% in pairwise comparisons, but they all share the same essential features of the globin fold. This family of structures is thus the prime example of a situation where natural selection has produced proteins whose amino acid sequences have diverged widely (although some homology is usually still recognizable) but whose three-dimensional structure has been essentially preserved.

Arthur Lesk and Cyrus Chothia at the MRC Laboratory of Molecular Biology in Cambridge, UK, compared the family of globin structures with the aim of answering two general questions: How can amino acid sequences that are very different form proteins that are very similar in their three-dimensional structure? What is the mechanism by which proteins adapt to mutations in the course of their evolution?

The hydrophobic interior is preserved

To answer the first question, Lesk and Chothia examined in detail residues at structurally equivalent positions that are involved in helix-heme contacts and in packing the α helices against each other. After comparing the nine globin structures then known, the 59 positions they found that fulfilled these criteria were divided into 31 positions buried in the interior of the protein and 28 in contact with the heme group. These positions are the principal determinants of both the function and the three-dimensional structure of the globin family.

One might expect these positions to exhibit a higher degree of amino acid conservation and hence sequence identity than the rest of the molecule. This is not, however, the case for distantly related molecules that have low sequence identity and derive from distantly related species. The sequence identity of these residues is no greater than in the rest of the

Petitioner Merck, Ex. 1014, p. 42

molecule. Since the important residues involved in packing the α helices are not conserved, we used to assume that the changes which have occurred compensate each other in size. This is not the case either. The volumes occupied by the 31 buried residues vary considerably between individual members. Thus neither conserved sequence nor size-compensatory mutations in the hydrophobic core are important factors in preserving three-dimensional structure during evolution. We now know that this is also true for other proteins, such as the immunoglobulins.

Lesk and Chothia did find, however, that there is a striking preferential conservation of the hydrophobic character of the amino acids at the 59 buried positions, but that no such conservation occurs at positions exposed on the surface of the molecule. With a few exceptions on the surface, hydrophobic residues have replaced hydrophilic ones and vice versa. However, the case of sickle-cell hemoglobin, which is described below, shows that a charge balance must be preserved to avoid hydrophobic patches on the surface. In summary, the evolutionary divergence of these nine globins has been constrained primarily by an almost absolute conservation of the hydrophobicity of the residues buried in the helix-to-helix and helix-to-heme contacts.

Helix movements accommodate interior side-chain mutations

Lesk and Chothia also found a simple answer to the question of how proteins adapt to changes in size of buried residues. The mode of packing the α helices is the same in all the globin structures: the same types of packing of ridges into grooves occur in corresponding α helices in all these structures. However, the relative positions and orientations of the α helices change to accommodate changes in the volume of side chains involved in the packing.

The proteins thus adapt to mutations of buried residues by changing their overall structure, which in the globins involves movements of entire α helices relative to each other. The structure of loop regions changes so that the movement of one α helix is not transmitted to the rest of the structure. Only movements that preserve the geometry of the heme pocket are accepted. Mutations that cause such structural shifts are tolerated because many different combinations of side chains can produce well-packed helix-helix interfaces of similar but not identical geometry and because the shifts are coupled so that the geometry of the active site is retained.

Sickle-cell hemoglobin confers resistance to malaria

Sickle-cell anemia is the classic example of an inherited disease that is caused by a change in a protein's amino acid sequence. Linus Pauling proposed in 1949 that it was caused by a defect in the hemoglobin molecule; he thus coined the term **molecular disease.** Seven years later Vernon Ingram showed that the disease was caused by a single mutation, a change in residue 6 of the β chain of hemoglobin from Glu to Val.

Hemoglobin is a tetramer built up of two copies each of two different polypeptide chains, α- and β-globin chains in normal adults. Each of the four chains has the globin fold with a heme pocket. Residue 6 in the β chain is on the surface of α helix A, and it is also on the surface of the tetrameric molecule (Figure 3.13).

The hemoglobin concentration in red blood cells, erythrocytes, is extremely high, 340 mg/ml. This is almost as high as in the crystalline state: the hemoglobin molecules, which are spheroids of dimension $50 \times 55 \times 65$ Å, are on average only 10 Å apart in the cells. It is thus surprising that they can nevertheless rotate and flow past one another. The mutation in sickle-cell hemoglobin converts a charged residue to a hydrophobic residue, and as a result, it produces a hydrophobic patch on the surface. This patch happens to fit and bind a hydrophobic pocket in the deoxygenated form of another

Petitioner Merck, Ex. 1014, p. 43

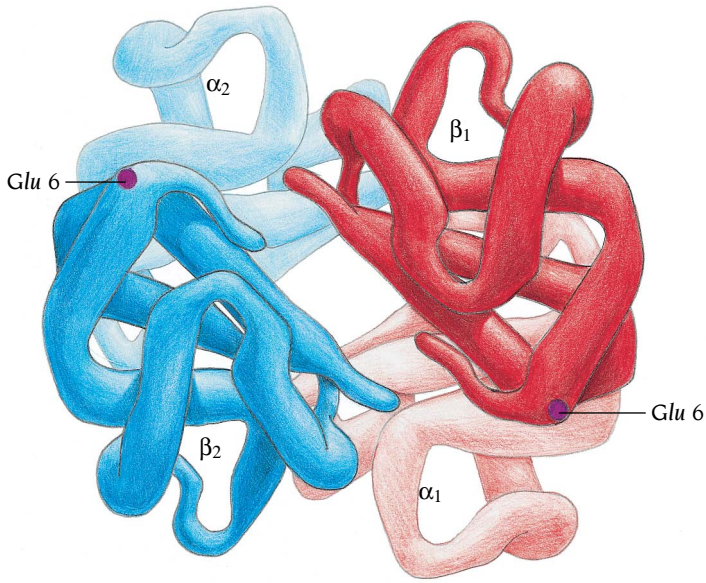


Figure 3.13 The hemoglobin molecule is built up of four polypeptide chains: two α chains and two β chains. Compare this with Figure 1.1 and note that for purposes of clarity parts of the α chains are not shown here. Each chain has a three-dimensional structure similar to that of myoglobin: the globin fold. In sickle-cell hemoglobin Glu 6 in the β chain is mutated to Val, thereby creating a hydrophobic patch on the surface of the molecule. The structure of hemoglobin was determined in 1968 to 2.8 Å resolution in the laboratory of Max Perutz at the MRC Laboratory of Molecular Biology, Cambridge, UK.

hemoglobin molecule (Figure 3.14a). In the oxygenated form of hemoglobin the shape of this pocket is slightly different, and thus no interaction occurs between hemoglobin molecules in the lungs. However, when hemoglobin in the blood capillaries delivers its oxygen, these hydrophobic interactions can occur and the highly concentrated hemoglobin in the cells polymerizes into fibers (Figure 3.14b and c). These fibers stiffen the erythrocytes and deform them into a sickle shape: hence the name sickle-cell anemia. In heterozygotes, where only one β-globin allele is mutated, this occurs only to a minor extent, whereas it is lethal for homozygotes, all of whose hemoglobin molecules carry the mutation.

We thus have here a case where a mutation on the surface of the globin fold, replacing a hydrophilic residue with a hydrophobic one, changes important properties of the molecule and produces a lethal disease. Why has the

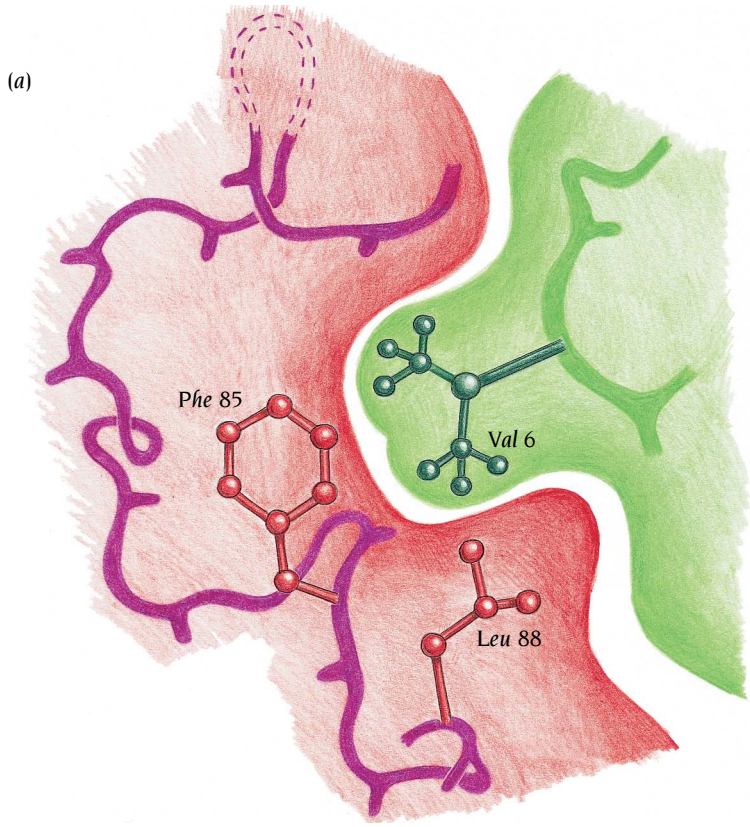
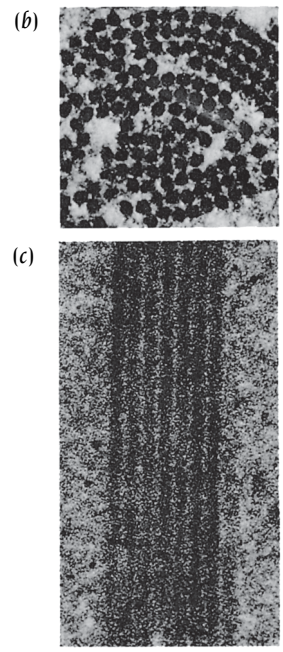


Figure 3.14 Sickle-cell hemoglobin molecules polymerize due to the hydrophobic patch introduced by the mutation Glu 6 to Val in the β chain. The diagram (a) illustrates how this hydrophobic patch (green) interacts with a hydrophobic pocket (red) in a second hemoglobin molecule, whose hydrophobic patch interacts with the pocket in a third molecule, and so on. Electron micrographs of sickle-cell hemoglobin fibers are shown in cross-section in (b) and along the fibers in (c). [(b) and (c) from J.T. Finch et al., *Proc. Natl. Acad. Sci. USA* 70: 718–722, 1973.]



Petitioner Merck, Ex. 1014, p. 44

mutation survived during evolution? It turns out that the disease gives increased resistance to malaria which has had a high survival value for heterozygotes, especially in Africa. In evolutionary terms, the death of homozygotes has been an acceptable price to pay for increased survival of heterozygotes in a malarial environment.

Conclusion

Coiled-coil α-helical structures are found both in fibrous proteins and as parts of smaller domains in many globular proteins. Alpha- (α-) domain structures consist of a bundle of α helices that are packed together to form a hydrophobic core. A common motif is the four-helix bundle structure, where four helices are pairwise arranged in either a parallel or an antiparallel fashion and packed against each other. The most intensively studied α structure is the globin fold, which has been found in a large group of related proteins, including myoglobin and hemoglobin. This structure comprises eight α helices that wrap around the core in different directions and form a pocket where the heme group is bound.

Rules have been derived that explain the different geometrical arrangements of α helices observed in α-domain structures. The helix packing in coiled-coil structures is determined by fitting the knobs of side chains in the first helix into holes between side chains in the second helix. For other α-helical structures the helix packing is determined by fitting ridges of side chains along one α helix into grooves between side chains of another helix.

The globin fold has been used to study evolutionary constraints for maintaining structure and function. Evolutionary divergence is primarily constrained by conservation of the hydrophobicity of buried residues. In contrast, neither conserved sequence nor size-compensatory mutations in the hydrophobic core are important. Proteins adapt to mutations in buried residues by small changes of overall structure that in the globins involve movements of entire helices relative to each other.

Selected readings

General

Chothia, C., Lesk, A.M. Helix movements in proteins. *Trends Biochem. Sci.* 13: 116–118, 1985.

Chothia, C., Levitt, M., Richardson, D. Helix-to-helix packing in proteins. *J. Mol. Biol.* 145: 215–250, 1981.

Dickerson, R.E., Geis, I. *Hemoglobin: Structure, Function, Evolution and Pathology*. Menlo Park, CA: Benjamin/Cummings, 1983.

Lesk, A.M., Chothia, C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* 136: 225–270, 1980.

Murzin, A.G., Finkelstein, A.V. General architecture of the α-helical globule. *J. Mol. Biol.* 204: 749–769, 1988.

Pauling, L., et al. Sickle cell anemia: a molecular disease. *Science* 110: 543–548, 1949.

Perutz, M.F. Hemoglobin structure and respiratory transport. *Sci. Am.* 239(6): 92–125, 1978.

Perutz, M.F. *Protein Structure: New Approaches to Disease and Therapy.* New York: Freeman, 1992.

Specific structures

Argos, P., Rossmann, M.G., Johnsson, J.E. A four-helical super-secondary structure. *Biochem. Biophys. Res. Comm.* 75: 83–86, 1977.

Banner, D.W., Kokkinidis, M., Tsernoglou, D. Structure of the col1 rop protein at 1.7 Å resolution. *J. Mol. Biol.* 196: 657–675, 1987.

Bashford, D., Chothia, C., Lesk, A.M. Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.* 196: 199–216, 1987.

Bloomer, A.C., et al. Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between subunits. *Nature* 276: 362–368, 1978.

Blundell, T., et al. Solvent-induced distortions and the curvature of α-helices. *Nature* 306: 281–283, 1983.

Clegg, G.A., et al. Helix packing and subunit conformation in horse spleen apoferritin. *Nature* 288: 298–300, 1980.

Cohen, C., Parry, D.A.D. Alpha-helical coiled coils—a widespread motif in proteins. *Trends Biochem. Sci.* 11: 245–248, 1986.

Crick, F.H.C. The packing of α-helices: simple coiled coils. *Acta Cryst.* 6: 689–697, 1953.

Petitioner Merck, Ex. 1014, p. 45

de Vos, A.M., Ultsch, M., Kossiakoff, A.A. Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science* 255: 306–312, 1992.

Embury, S.H. The clinical pathophysiology of sickle-cell disease. *Annu. Rev. Med.* 37: 361–376, 1986.

Fermi, G., et al. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* 175: 159–174, 1984.

Fermi, G., Perutz, M.F. *Atlas of Molecular Structures in Biology. 2. Haemoglobin and Myoglobin.* Oxford, UK: Clarendon Press, 1981.

Finch, J.T., et al. Structure of sickled erythrocytes and of sickle-cell hemoglobin fibers. *Proc. Natl. Acad. Sci. USA* 70: 718–722, 1973.

Finzel, B.C., et al. Structure of ferricytochrome c′ from *Rhodospirillum molischianum* at 1.67 Å resolution. *J. Mol. Biol.* 186: 627–643, 1985.

Ingram, V.M. Gene mutation in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* 180: 326–328, 1957.

Lederer, F., et al. Improvement of the 2.5 Å resolution model of cytochrome b_{562} by redetermining the primary structure and using molecular graphics. *J. Mol. Biol.* 148: 427–448, 1981.

Nordlund, P., Sjoberg, B.-M., Eklund, H. Three-dimensional structure of the free radical protein of ribonucleotide reductase. *Nature* 345: 593–598, 1990.

Pastore, A., et al. Structural alignment and analysis of two distantly related proteins: *Aplysia limacina* myoglobin and sea lamprey globin. *Proteins* 4: 240–250, 1988.

Pastore, A., Lesk, A.M. Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins* 8: 133–155, 1990.

Phillips, S.E.V. Structure and refinement of oxymyoglobin at 1.6 Å resolution. *J. Mol. Biol.* 142: 531–554, 1980.

Presnell, S.R., Cohen, F.E. Topological distribution of four-α-helix bundles. *Proc. Natl. Acad. Sci. USA* 86: 6592–6596, 1989.

Richmond, T.J., Richards, F.M. Packing of α-helices: geometrical constraints and contact areas. *J. Mol. Biol.* 119: 537–555, 1978.

Sheriff, S., Hendrickson, W.A., Smith, J.L. Structure of myohemerythrin in the azidomet state at 1.7/1.3 Å resolution. *J. Mol. Biol.* 197: 273–296, 1987.

Thunissen, A.-M., et al. Doughnut-shaped structure of a bacterial muramidase revealed by x-ray crystallography. *Nature* 367: 750–753, 1994.

Watson, H.C. The stereochemistry of the protein myoglobin. *Progr. Stereochem.* 4: 299–333, 1969.

Weber, P.C., Salemme, F.R. Structural and functional diversity in 4-α-helical proteins. *Nature* 287: 82–84, 1980.

Petitioner Merck, Ex. 1014, p. 46

Alpha/Beta Structures

4

The most frequent of the domain structures are the alpha/beta (α/β) domains, which consist of a central parallel or mixed β sheet surrounded by α helices. All the glycolytic enzymes are α/β structures as are many other enzymes as well as proteins that bind and transport metabolites. In α/β domains, binding crevices are formed by loop regions. These regions do not contribute to the structural stability of the fold but participate in binding and catalytic action.

Parallel β strands are arranged in barrels or sheets

There are three main classes of α/β proteins. In the first class there is a core of twisted parallel β strands arranged close together, like the staves of a barrel. The α helices that connect the parallel β strands are on the outside of this barrel (Figure 4.1a). This domain structure is often called the **TIM barrel** from the structure of the enzyme triosephosphate isomerase, where it was first observed. The second class contains an open twisted β sheet surrounded by α helices on both sides. A typical example is shown in Figure 4.1b, a nucleotide-binding domain sometimes called the **Rossman fold** after Michael Rossman, Purdue University, who first discovered this fold in the enzyme lactate dehydrogenase in 1970. The third class is formed by amino acid sequences that contain repetitive regions of a specific pattern of leucine residues, so-called **leucine-rich motifs**, which form α helices and β strands. The β strands form a curved parallel β sheet with all the α helices on the outside. The structure of one member of this class, a ribonuclease inhibitor (illustrated in Figure 4.11), is shaped like a horseshoe, and consequently this class is called the **horseshoe fold.**

Barrels, open sheets, and horseshoe structures are all built up from β-α-β motifs. To illustrate how they differ, let us consider two β-α-β motifs: β_1-$\alpha_{1,2}$-β_2 and β_3-$\alpha_{3,4}$-β_4 linked together by helix $\alpha_{2,3}$. There are two fundamentally different ways these two motifs can be connected into a β sheet of four parallel strands, as shown in Figure 4.2. Strand β_3 can be aligned adjacent either to strand β_2, giving the strand order 1 2 3 4, or to strand β_1, giving the strand order 4 3 1 2. In the first case the two β-α-β motifs are joined with the same orientation. Since the β-α-β unit is almost always a right-handed structure, all three α helices (one from each motif and the joining helix) are on the same side, above the β sheet (Figure 4.2a). In barrel and horseshoe structures the β-α-β motifs are linked in this way and consist of consecutive β-α-β units, all in the same orientation.

Petitioner Merck, Ex. 1014, p. 47

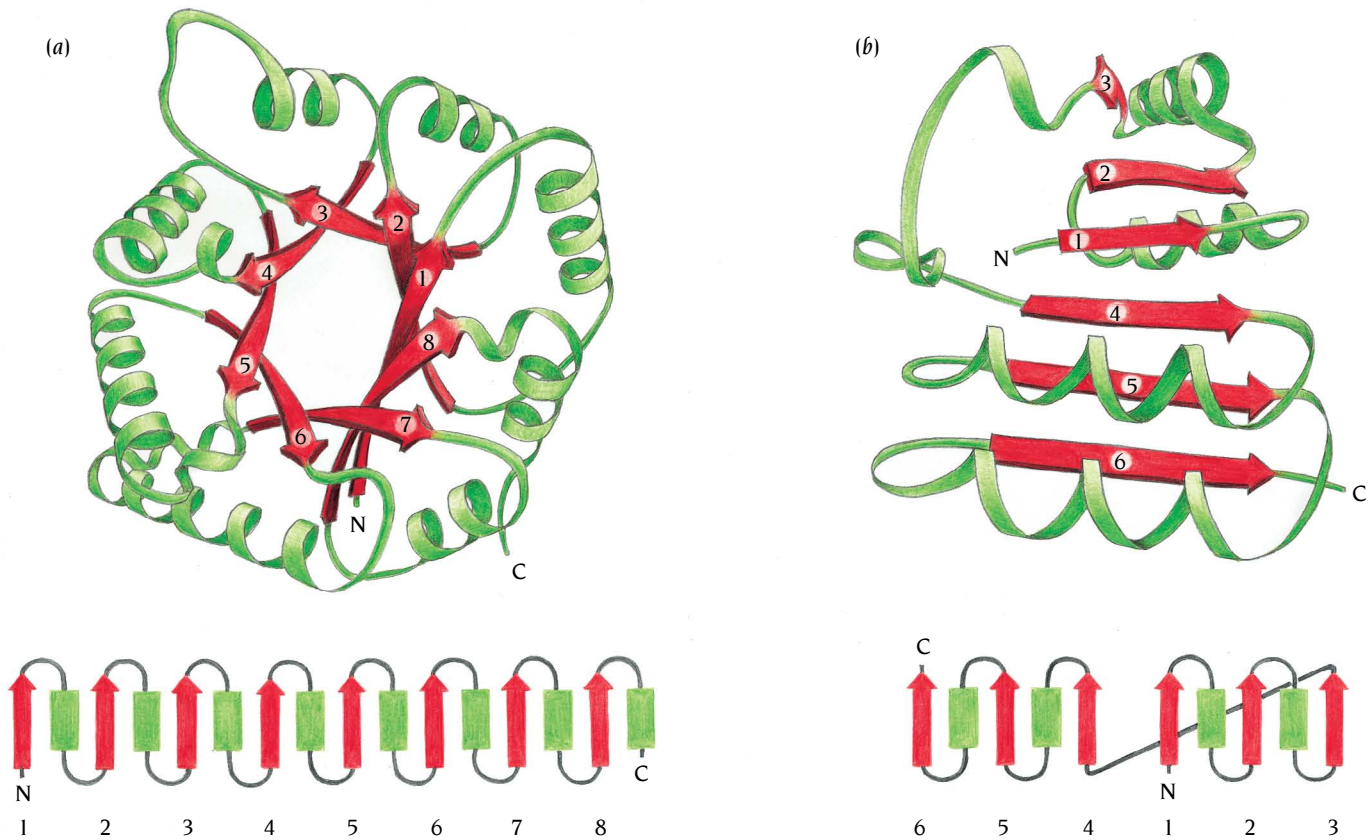


Figure 4.1 Alpha/beta domains are found in many proteins. They occur in different classes, two of which are shown here: (a) a closed barrel exemplified by schematic and topological diagrams of the enzyme triosephosphate isomerase and (b) an open twisted sheet with helices on both sides, as in the coenzyme-binding domain of some dehydrogenases. Both classes are built up from β-α-β motifs that are linked such that the β strands are parallel. Rectangles represent α helices, and arrows represent β strands in the topological diagrams. [(a) Adapted from J. Richardson. (b) Adapted from B. Furugren.]

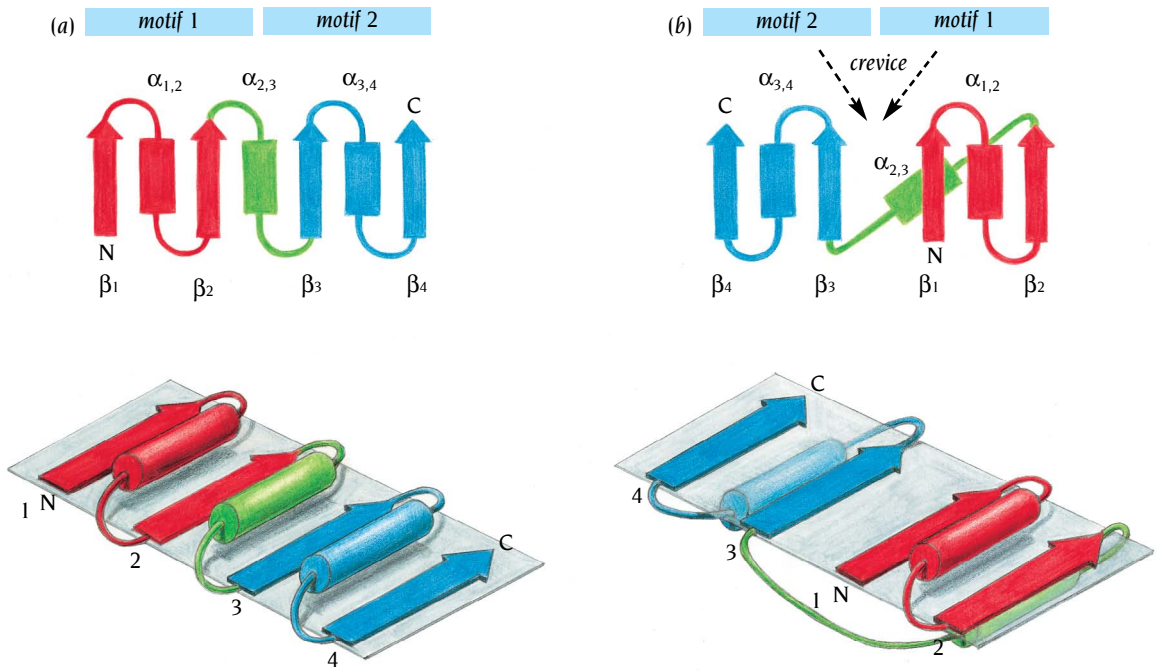
In the second case we must turn the second motif around in order to align β strands 1 and 3. As a result of the right-handed structure of the β-α-β motif, its α helix is on the other side of the β sheet (Figure 4.2b). In open twisted β-sheet structures there are always one or more such alignments and therefore there are α helices on both sides of the β sheet. These geometric rules apply because virtually all β-α-β motifs are right-handed. As pointed out in Chapter 2, this is an empirical rule that almost always applies, although no convincing explanation has been found. `

Alpha/beta barrels occur in many different enzymes

In α/β structures where the strand order is 1 2 3 4, all connections are on the same side of the β sheet. An open twisted β sheet of this sort with four or more parallel β strands would leave one side of the parallel β sheet exposed to the solvent and the other side shielded by the α helices. Such a domain structure is rarely observed, except in the horseshoe structure or as part of more complex structures where loop regions, extra α helices, or additional β sheets cover the exposed side of the β sheet. Instead, a closed barrel of twisted β strands is formed with all the connecting α helices on the outside of the barrel, as shown in Figure 4.1a. However, more than four β strands are needed to provide enough staves to form a closed barrel, and almost all the closed α/β barrels observed to date have eight parallel β strands. These are arranged such that β strand 8 is adjacent and hydrogen-bonded to β strand 1. In a few cases the barrels do not have eight parallel β strands; there are also barrels that contain ten parallel β strands and some that contain eight parallel and two antiparallel β strands. In almost all cases the cross-connections between the parallel β strands are α helices; in addition, there is usually an α helix after the last β strand.

The eight-stranded α/β-barrel structure is one of the largest and most regular of all domain structures. A minimum of about 200 residues are required to form this structure. It has been found in many different proteins, most of which are enzymes, with completely different amino acid sequences and

Petitioner Merck, Ex. 1014, p. 48



different functions. Superimposing the structures of these proteins shows that around 160 residues are structurally equivalent. These residues form the β strands and α helices. The remaining residues form the loop regions that connect the β strands with the α helices. These loops have quite different lengths and conformations in the different proteins. This reflects the fact that the β strands and α helices form the structural framework of the enzyme, whereas the loops contain the amino acids responsible for its catalytic chemistry. In some cases the loops are very long and form independent domains in the overall subunit structure.

Figure 4.2 A β-α-β motif is a right-handed structure. Two such motifs can be joined into a four-stranded parallel β sheet in two different ways. They can be aligned with the α helices either on the same side of the β sheet (a) or on opposite sides (b). In case (a) the last β strand of motif 1 (red) is adjacent to the first β strand of motif 2 (blue), giving the strand order 1 2 3 4. The motifs are aligned in this way in barrel structures (see Figure 4.1a) and in the horseshoe fold (see Figure 4.11). In case (b) the first β strands of both motifs are adjacent, giving the strand order 4 3 1 2. Open twisted sheets (see Figure 4.1b) contain at least one motif alignment of this kind. In both cases the motifs are joined by an α helix (green).

Branched hydrophobic side chains dominate the core of α/β barrels

In barrels the hydrophobic side chains of the α helices are packed against hydrophobic side chains of the β sheet. The α helices are antiparallel and adjacent to the β strands that they connect. Thus the barrel is provided with a shell of hydrophobic residues from the α helices and the β strands.

Since the side chains of consecutive amino acids of a β strand are on opposite sides of the β sheet, every second residue of the β strands contributes to this hydrophobic shell. The other side chains of the β strands point inside the barrel to form a hydrophobic core; this core is therefore comprised exclusively of side chains of β-strand residues (Figure 4.3).

The packing interactions between α helices and β strands are dominated by the residues Val (V), Ile (I), and Leu (L), which have branched hydrophobic side chains. This is reflected in the amino acid composition: these three amino acids comprise approximately 40% of the residues of the β strands in parallel β sheets. The important role that these residues play in packing α helices against β sheets is particularly obvious in α/β-barrel structures, as shown in Table 4.1.

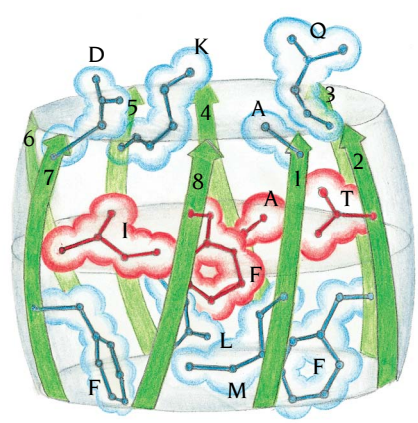


Figure 4.3 In most α/β-barrel structures the eight β strands of the barrel enclose a tightly packed hydrophobic core formed entirely by side chains from the β strands. The core is arranged in three layers, with each layer containing four side chains from alternate β strands. The schematic diagram shows this packing arrangement in the α/β barrel of the enzyme glycolate oxidase, the structure of which was determined by Carl Branden and colleagues in Uppsala, Sweden.

Petitioner Merck, Ex. 1014, p. 49

Table 4.1 The amino acid residues of the eight parallel β strands in the barrel structure of the enzyme triosephosphate isomerase from chicken muscle

		Positions				
Strand no.	Residue no.	1	2	3	4	5
1	6–10	Phe	Val	Gly	Gly	Asn
2	37–41	Glu	Val	Val	Cys	Gly
3	59–63	Gly	Val	Ala	Ala	Gln
4	89–93	Trp	Val	Ile	Leu	Gly
5	121–125	Gly	Val	Ile	Ala	Cys
6	158–162	Lys	Val	Val	Leu	Ala
7	204–208	Arg	Ile	Ile	Tyr	Gly
8	227–231	Gly	Phe	Leu	Val	Gly

The sequences are aligned so that residues in positions 1, 3, and 5 point into the barrel and residues in positions 2 and 4 point toward the α helices on the outside and are involved in the hydrophobic interactions between the β strands and the α helices.

Bulky hydrophobic residues from positions 1, 3, and 5 of the β strands fill the interior of the barrel and form a tightly packed hydrophobic core (see Figure 4.3). Note from Table 4.1 that some of these residues are Lys, Arg, or Gln, which have a polar end group (see Panel 1.1, pp. 6–7) terminating a chain of hydrophobic $-CH_2$ groups. These chains are in the hydrophobic interior and traverse part of the barrel; their polar end groups are on the top or bottom surface of the barrel and are in contact with the aqueous environment. By this arrangement even amino acids that are classified as polar can participate in the formation of hydrophobic cores of compact globular domains through the hydrophobic parts of their side chains.

There is one exception to the rule that requires bulky hydrophobic residues to fill the interior of eight-stranded α/β barrels in order to form a tightly packed hydrophobic core. The coenzyme B_{12}–dependent enzyme methylmalonyl–coenzyme A mutase, the x-ray structure of which was determined by Phil Evans and colleagues at the MRC Laboratory of Molecular

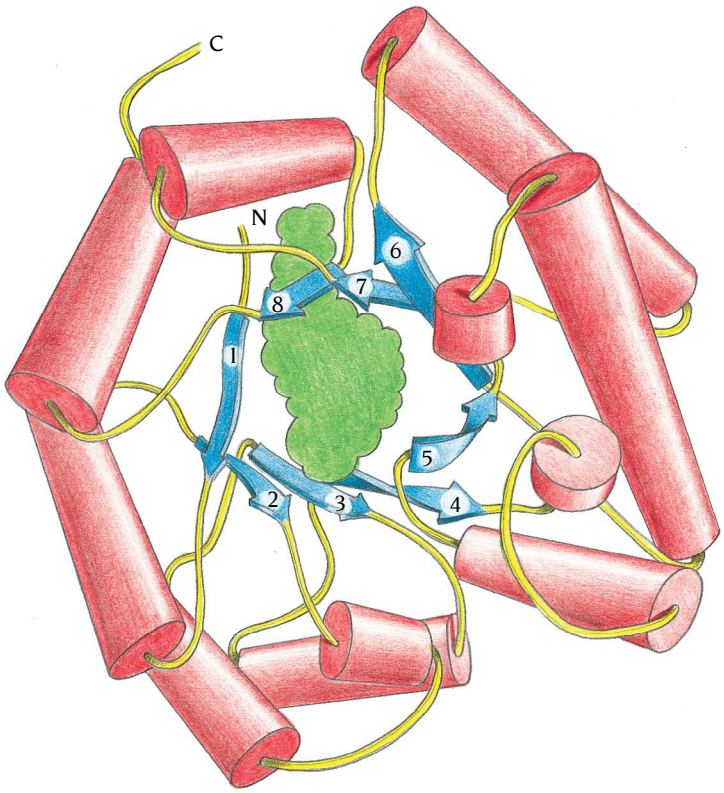


Figure 4.4 Schematic diagram of the structure of the α/β-barrel domain of the enzyme methylmalonyl–coenzyme A mutase. Alpha helices are red, and β strands are blue. The inside of the barrel is lined by small hydrophilic side chains (serine and threonine) from the β strands, which creates a hole in the middle where one of the substrate molecules, coenzyme A (green), binds along the axis of the barrel from one end to the other. (Adapted from a computer-generated diagram provided by P. Evans.)

Petitioner Merck, Ex. 1014, p. 50

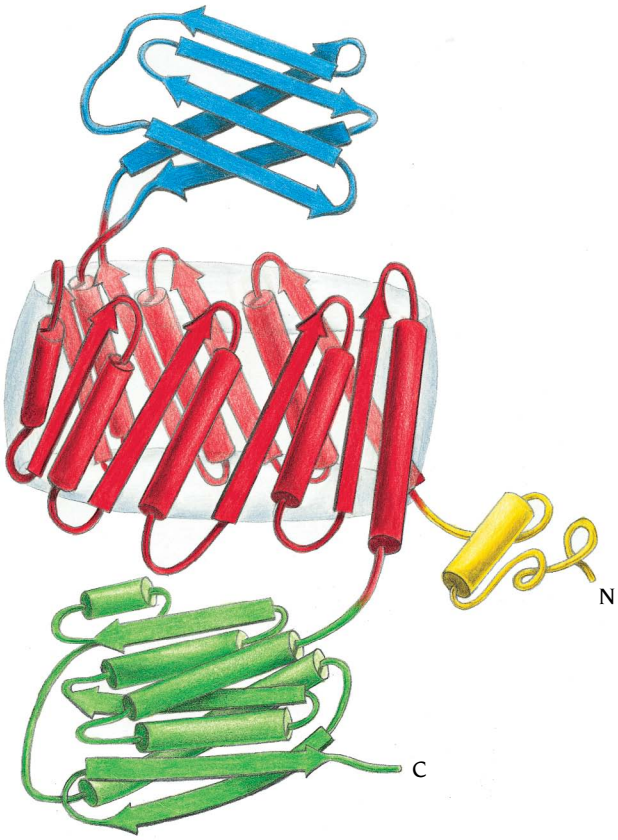


Figure 4.5 The polypeptide chain of the enzyme pyruvate kinase folds into several domains, one of which is an α/β barrel (red). One of the loop regions in this barrel domain is extended and comprises about 100 amino acid residues that fold into a separate domain (blue) built up from antiparallel β strands. The C-terminal region of about 140 residues forms a third domain (green), which is an open twisted α/β structure.

Biology, in Cambridge, UK, has a hole in the middle of its α/β-barrel domain (Figure 4.4). The serine and threonine side chains that project from the β strands into the interior of the barrel are small and polar, and therefore do not fill up the space available inside the barrel. The resulting tunnel through the barrel provides an ideal environment for the catalytic reaction and is sufficiently large for the substrate molecule, methylmalonyl–coenzyme A, to bind. Many enzyme-catalyzed reactions, including the reaction catalyzed by this mutase, require that the reactive part of the substrate molecule be shielded from solvent during the catalytic reaction. When the substrate is bound in the tunnel inside the barrel, it is shielded from the outside world. In α/β barrels with a hydrophobic core, the substrate binds at the surface of the barrel and conformational changes of loop regions shield the substrate from the solvent.

Pyruvate kinase contains several domains, one of which is an α/β barrel

All known eight-stranded α/β-barrel domains have enzymatic functions that include isomerization of small sugar molecules, oxidation by flavin coenzymes, phosphate transfer, and degradation of sugar polymers. In some of these enzymes the barrel domain comprises the whole subunit of the protein; in others the polypeptide chain is longer and forms several additional domains. An enzymatic function in these multidomain subunits, however, is always associated with the barrel domain.

For example, each subunit of the dimeric glycolytic enzyme triosephosphate isomerase (see Figure 4.1a) consists of one such barrel domain. The polypeptide chain has 248 residues in which the first β strand of the barrel starts at residue 6 and the last α helix of the barrel ends at residue 246. In contrast, the subunit of the glycolytic enzyme pyruvate kinase (Figure 4.5), which was solved at 2.6 Å resolution in the laboratory of Hilary Muirhead, Bristol University, UK, is folded into four different domains. The polypeptide chain of this cat muscle enzyme has 530 residues. In Figure 4.5, residues 1–42

Petitioner Merck, Ex. 1014, p. 51

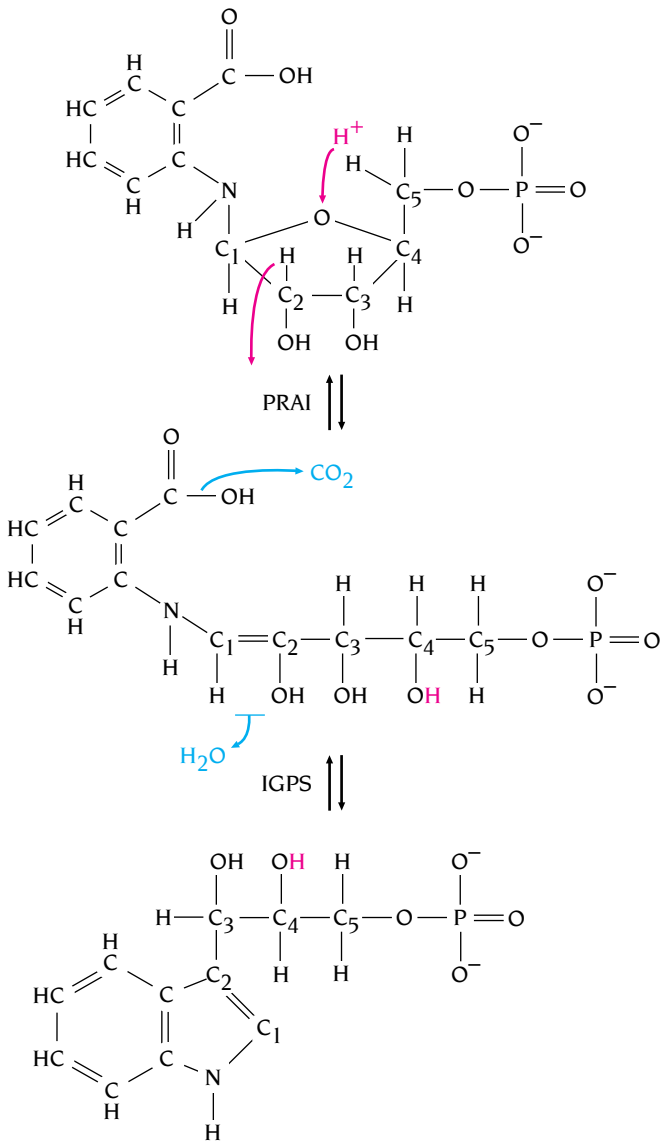


Figure 4.6 The bifunctional enzyme PRA-isomerase (PRAI):IGP-synthase (IGPS) catalyzes two sequential reactions in the biosynthesis of tryptophan. In the first reaction (top half), which is catalyzed by the C-terminal PRAI domain of the enzyme, the substrate N-(5′-phosphoribosyl) anthranilate (PRA) is converted to 1-(*o*-carboxyphenylamino)-1-deoxyribulose 5-phosphate (CdRP) by a rearrangement reaction. The succeeding step (bottom half), a ring closure reaction from CdRP to indole-3-glycerol phosphate (IGP), is catalyzed by the N-terminal IGPS domain.

form a small domain (yellow) involved in subunit contacts in the tetrameric molecule; residues 43–115 and 224–387 form an α/β-barrel domain (red) that binds substrate and provides the catalytic groups; residues 116–223 loop out from the end of β-strand number 3 in the barrel domain and are folded into a separate domain consisting of an antiparallel β sheet (blue); and finally, residues 388–530 form an open twisted α/β domain (green). This structure illustrates perfectly how a long polypeptide chain can be arranged in domains of different structural types.

Double barrels have occurred by gene fusion

PRA-isomerase:IGP-synthase, a bifunctional enzyme from *E. coli* that catalyzes two reactions in the synthesis of tryptophan (Figure 4.6), has a polypeptide chain that forms two α/β barrels. The structure of this enzyme, solved at 2.8 Å in the laboratory of Hans Jansonius in Basel, Switzerland, showed that residues 48–254 form one barrel with IGP-synthase activity, while residues 255–450 form the second barrel with PRA-isomerase activity (Figure 4.7).

In *Bacillus subtilis* these two reactions are catalyzed by two separate enzymes that have amino acid sequences homologous to the corresponding regions of the bifunctional enzyme from *E. coli,* and thus each forms a barrel

Petitioner Merck, Ex. 1014, p. 52

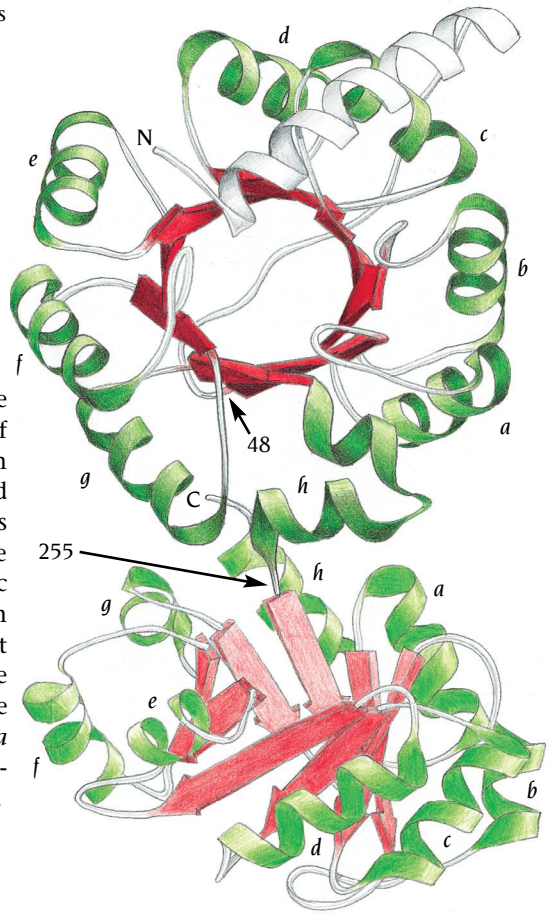


Figure 4.7 Two of the enzymatic activities involved in the biosynthesis of tryptophan in *E. coli,* phosphoribosyl anthranilate (PRA) isomerase and indoleglycerol phosphate (IGP) synthase, are performed by two separate domains in the polypeptide chain of a bifunctional enzyme. Both these domains are α/β-barrel structures, oriented such that their active sites are on opposite sides of the molecule. The two catalytic reactions are therefore independent of each other. The diagram shows the IGP-synthase domain (residues 48–254) with dark colors and the PRA-isomerase domain with light colors. The α helices are sequentially labeled a–h in both barrel domains. Residue 255 (arrow) is the first residue of the second domain. (Adapted from J.P. Priestle et al., *Proc. Natl. Acad. Sci. USA* 84: 5690–5694, 1987.)

structure. There is no obvious functional advantage to *E. coli* in having these two enzymatic activities in one polypeptide chain, since the active sites of the two barrels are on opposite sides of the molecule facing away from each other and the two reactions are thus independent of each other. A third organism, *Neurospora crassa,* has an enzyme with three catalytic activities within the same polypeptide chain; here two domains similar to those of the *E. coli* enzyme are linked to a third domain that has yet another enzymatic function in the same biosynthetic pathway. These differences between species reflect different ways to organize the genome. DNA sequences that code for protein domains with different functions are organized into separate genes in one organism and fused into a single gene in another. Although the three-dimensional structures of these enzymes in *B. subtilis* and *N. crassa* have not been solved by crystallography, we can be certain that they are α/β-barrel domains because of their sequence homologies to the *E. coli* proteins.

The active site is formed by loops at one end of the α/β barrel

In all these α/β-barrel domains the active site is in a very similar position. It is situated in the bottom of a funnel-shaped pocket created by the eight loops that connect the carboxy end of the β strands with the amino end of the α helices (Figure 4.8). Residues that participate in binding and catalysis are in

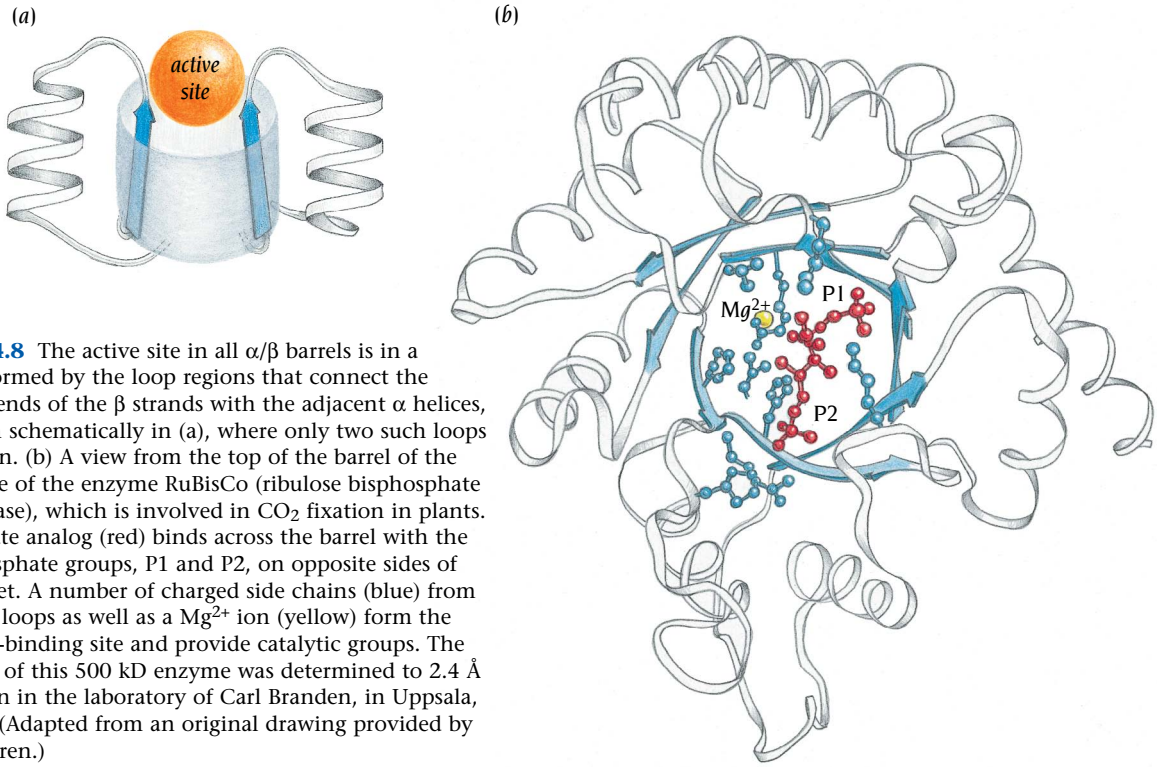


Figure 4.8 The active site in all α/β barrels is in a pocket formed by the loop regions that connect the carboxy ends of the β strands with the adjacent α helices, as shown schematically in (a), where only two such loops are shown. (b) A view from the top of the barrel of the active site of the enzyme RuBisCo (ribulose bisphosphate carboxylase), which is involved in CO_2 fixation in plants. A substrate analog (red) binds across the barrel with the two phosphate groups, P1 and P2, on opposite sides of the pocket. A number of charged side chains (blue) from different loops as well as a Mg^{2+} ion (yellow) form the substrate-binding site and provide catalytic groups. The structure of this 500 kD enzyme was determined to 2.4 Å resolution in the laboratory of Carl Branden, in Uppsala, Sweden. (Adapted from an original drawing provided by Bo Furugren.)

Petitioner Merck, Ex. 1014, p. 53

these loop regions. In other words, these enzymes are modeled on a common stable scaffold of eight parallel β strands surrounded by eight α helices. In each case the specific enzymatic activity is determined by the eight loop regions at the carboxy end of the β strands, which do not contribute to the structural stability of the scaffold. In some cases an additional loop region from a second domain or a different subunit comes close to this active site and also participates in binding and catalysis.

Alpha/beta barrels provide examples of evolution of new enzyme activities

How do new enzyme activities evolve? Are new enzymes formed from random sequences generated by recombination and other genetic rearrangements or do they arise by divergent evolution from a preexisting set of enzymes. Greg Petsko at Brandeis University has provided strong evidence for the latter case from studies of α/β-barrel enzymes in a rare metabolic pathway, conversion of mandelate to benzoate. This rare metabolic pathway is thought to be of recent evolutionary origin, since it is present in only a few pseudomonad species.

The first enzyme in this pathway, mandelate racemase, catalyzes the interconversion of the two optical isomers of mandelate (Figure 4.9a). The key step in this reaction is proton abstraction from a carbon atom, producing an enolic intermediate. Petsko found that the three-dimensional structure of this enzyme, including its α/β barrel, is very similar to that of a quite different enzyme, muconate lactonizing enzyme, which catalyzes a different chemical reaction (Figure 4.9b) but which also involves the formation of an intermediate by proton abstraction. The amino acid sequences of the 350 residues of these enzymes showed 26% sequence identity, which clearly demonstrates that they are evolutionarily related. By comparing these two structures in detail Petsko found significant similarities in the region of the active site that catalyzes proton abstraction and intermediate formation but substantial differences in those regions of the active site that confer substrate specificity.

These results are compatible with an evolutionary history in which the new enzyme activity of mandelate racemase has evolved from a preexisting enzyme that catalyzes the basic chemical reaction of proton abstraction and formation of an intermediate. Subsequent mutations have modified the

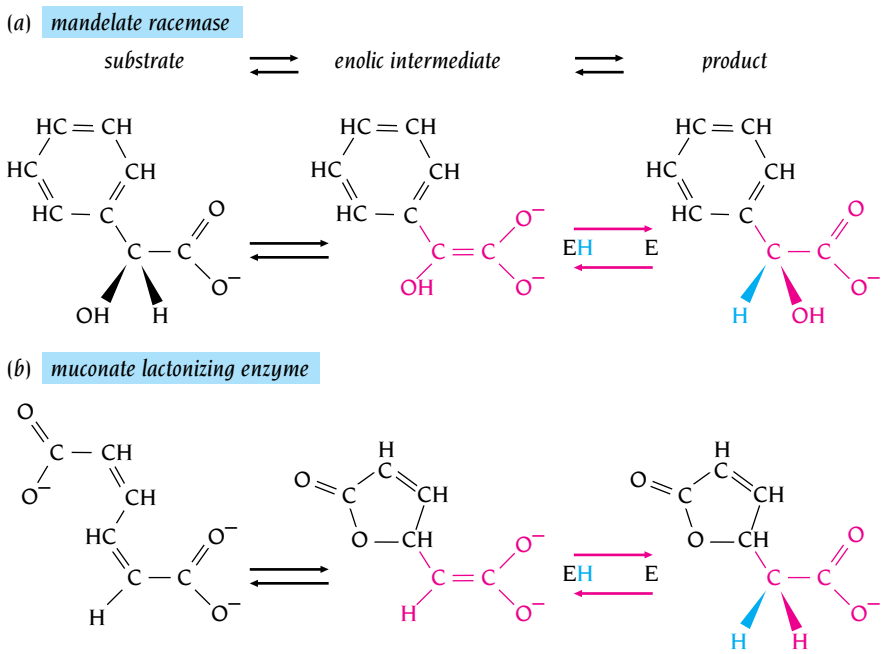


Figure 4.9 Mechanisms of the reactions catalyzed by the enzymes mandelate racemase (a) and muconate lactonizing enzyme (b). The two overall reactions are quite different: a change of configuration of a carbon atom for mandelate racemase versus ring closure for the lactonizing enzyme. However, one crucial step (pink) in the two reactions is the same: addition of a proton (blue) to an intermediate of the substrate (pink) from a lysine residue of the enzyme (E) or, in the reverse direction, formation of an intermediate by proton abstraction from the carbon atom adjacent to the carboxylate group.

Petitioner Merck, Ex. 1014, p. 54

```
                  ←β strand→        loop        ←α helix→
                  2   5   7         12          17  20  24
(type A)  NH2  -X-L-E-X-L-X-L-X-X-C-X-L-T-X-X-X-C-X-X-L-X-X-a-L-X-X-X-X-

(type B)  NH2  -X-L-R-E-L-X-L-X-X-N-X-L-G-D-X-G-a-X-X-L-X-X-X-L-X-X-P-X-X-
```

Figure 4.10 Consensus amino acid sequence and secondary structure of the leucine-rich motifs of type A and type B. "X" denotes any amino acid; "a" denotes an aliphatic amino acid. Conserved residues are shown in bold in type B.

substrate specificity while preserving the ability to catalyze the basic chemical reaction. Chemistry is the important factor to preserve during evolution of new enzymes, while specificity can be modified. It would therefore seem that relatively nonspecific enzymes, which may have existed earlier in evolution or which may arise occasionally through random genetic rearrangements, are the clay from which nature sculpts new enzymes. To preserve the original enzyme activity and at the same time allow divergence, the precursor gene for the enzyme must first be duplicated at some point.

Further strong evidence that proteins with new functions evolve by the process of gene duplication and subsequent modification by mutation has recently been found by examination of the genome sequence of the bacterium *Haemophilus influenzae*, the first free-living organism to be completely sequenced. Sequence comparisons showed that, of the 1680 identified proteins coded by this genome, at least one third are related to one or more other proteins within the genome and therefore have arisen from processes that involve gene duplications.

Alpha/beta barrels are particularly well suited to such an evolutionary strategy, since substrate specificity and catalytic function reside in loop regions which are separated from the residues of the α helices and β strands that contribute to the structural stability of these domain structures. Such enzymes should also be excellent targets for genetic redesign *in vitro*. By changing the lengths and specific residues of the active-site loop regions, it might be possible to produce novel substrate specificities without affecting the stability of the structural framework and therefore the enzyme.

Leucine-rich motifs form an α/β-horseshoe fold

Leucine-rich motifs—tandem homologous amino acid sequences of about 20–30 residues—have been identified from sequence studies in over 60 different proteins, including receptors, cell adhesion molecules, bacterial virulence factors, and molecules involved in RNA splicing and DNA repair. The x-ray structure of one member of this class of proteins, a ribonuclease inhibitor, has been determined by Johann Deisenhofer and colleagues at the University of Texas, Dallas. The 456 amino acids of the polypeptide chain are arranged in 15 tandem leucine-rich motifs of two types that alternate along the chain: type A, with 29 residues, and type B, with 28 residues. In addition there are two short regions with nonhomologous sequences at the termini of the chain. The consensus sequence of these homologous repeats (Figure 4.10) indicates that both types of repeat contain a characteristic pattern of leucine residues that play an important structural role, as we will see.

Each repeat forms a right-handed β-loop-α structure similar to those found in the two other classes of α/β structures described earlier. Sequential β-loop-α repeats are joined together in a similar way to those in the α/β-barrel structures. The β strands form a parallel β sheet, and all the α helices are on one side of the β sheet. However, the β strands do not form a closed barrel; instead they form a curved open structure that resembles a horseshoe with α helices on the outside and a β sheet forming the inside wall of the horseshoe (Figure 4.11). One side of the β sheet faces the α helices and participates in a hydrophobic core between the α helices and the β sheet; the other side of the β sheet is exposed to solvent, a characteristic other α/β structures do not have.

The leucine residues in this leucine-rich motif form a hydrophobic core between the β sheet and the α helices. Leucine residues 2, 5, and 7 (see Figure

Petitioner Merck, Ex. 1014, p. 55

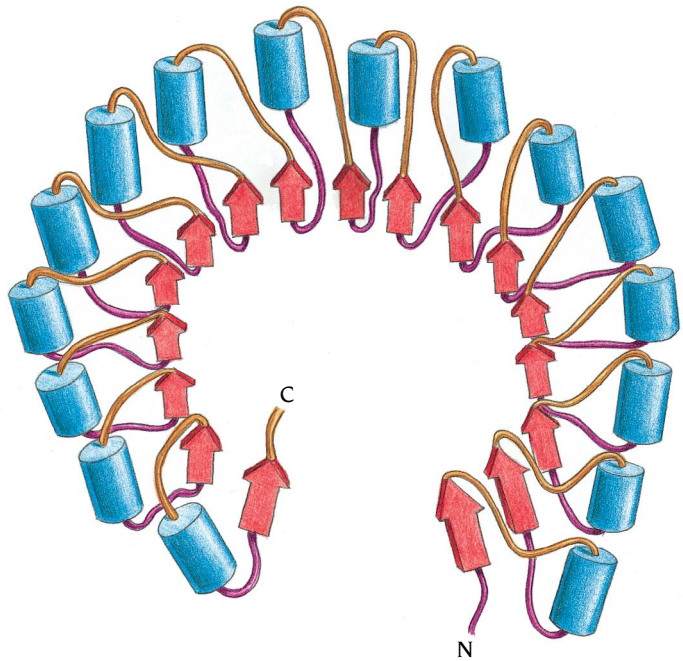


Figure 4.11 Schematic diagram of the structure of the ribonuclease inhibitor. The molecule, which is built up by repetitive β-loop-α motifs, resembles a horseshoe with a 17-stranded parallel β sheet on the inside and 16 α helices on the outside. The β sheet is light red, α helices are blue, and loops that are part of the β-loop-α motifs are orange. (Adapted from B. Kobe et al., *Nature* 366: 751–756, 1993.)

4.10) from the β strand of the motif pack against leucine residues 20 and 24 of the α helix to form the main part of the hydrophobic region. Leucine residue 12 from the loop region is also part of this hydrophobic core, as is residue 17 from the α helix, which is usually hydrophobic (Figure 4.12).

Leucine residues 2, 5, 7, 12, 20, and 24 of the motif are invariant in both type A and type B repeats of the ribonuclease inhibitor. An examination of more than 500 tandem repeats from 68 different proteins has shown that residues 20 and 24 can be other hydrophobic residues, whereas the remaining four leucine residues are present in all repeats. On the basis of the crystal structure of the ribonuclease inhibitor and the important structural role of these leucine residues, it has been possible to construct plausible structural models of several other proteins with leucine-rich motifs, such as the extracellular domains of the thyrotropin and gonadotropin receptors.

Alpha/beta twisted open-sheet structures contain α helices on both sides of the β sheet

In the next class of α/β structures there are α helices on both sides of the β sheet. This has at least three important consequences. First, a closed barrel cannot be formed unless the β strands completely enclose the α helices on one side of the β sheet. Such structures have never been found and are very unlikely to occur, since a large number of β strands would be required to enclose even a single α helix. Instead, the β strands are arranged into an open twisted β sheet such as that shown in Figure 4.1b.

Second, there are always two adjacent β strands (β_1 and β_3 in Figure 4.2b) in the interior of the β sheet whose connections to the flanking β strand are on opposite sides of the β sheet. One of the loops from one of these two β strands goes above the β sheet, whereas the other loop goes below. This creates a crevice outside the edge of the β sheet between these two loops (Figure 4.13). Almost all binding sites in this class of α/β proteins are located in crevices of this type at the carboxy edge of the β sheet, as we discuss in detail

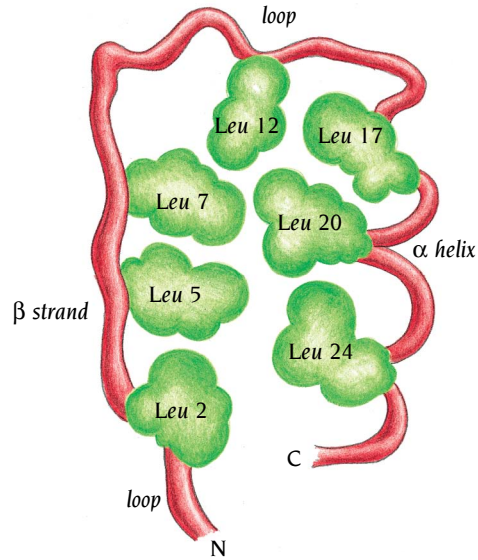


Figure 4.12 Schematic diagram illustrating the role of the conserved leucine residues (green) in the leucine-rich motif in stabilizing the β-loop-α structural module. In the ribonuclease inhibitor, leucine residues 2, 5, and 7 from the β strand pack against leucine residues 17, 20, and 24 from the α helix as well as leucine residue 12 from the loop to form a hydrophobic core between the β strand and the α helix.

Petitioner Merck, Ex. 1014, p. 56

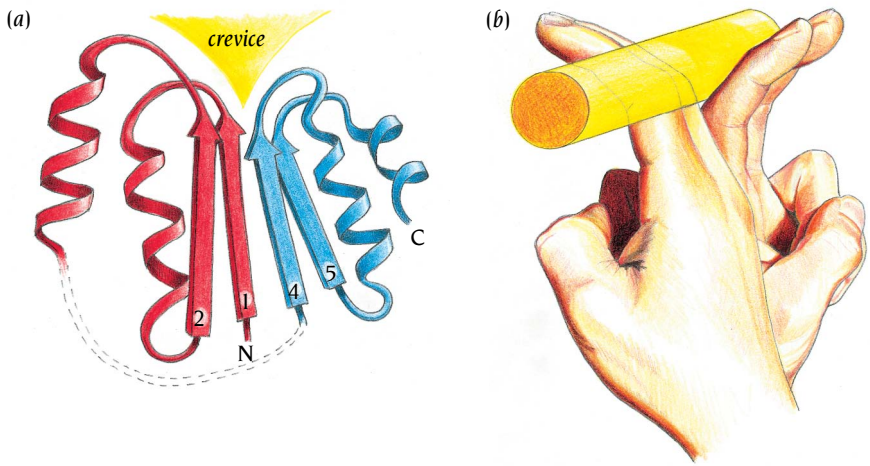


Figure 4.13 (a) The active site in open twisted α/β domains is in a crevice outside the carboxy ends of the β strands. This crevice is formed by two adjacent loop regions that connect the two strands with α helices on opposite sides of the β sheet. This is illustrated by the curled fingers of two hands (b), where the top halves of the fingers represent loop regions and the bottom halves represent the β strands. The rod represents a bound molecule in the binding crevice.

later. (We define the carboxy edge of the sheet as the edge that is formed by the carboxy ends of the parallel β strands in the sheet.)

Third, in open-sheet structures the α helices are packed against both sides of the β sheet. Each β strand thus contributes hydrophobic side chains to pack against α helices in two similar hydrophobic core regions, one on each side of the β sheet.

Open β-sheet structures have a variety of topologies

We have seen that all members of the α/β-barrel domain structures have the same basic arrangement of eight α helices and eight β strands. Within the open α/β sheets, however, there is much more variation in structure, as is obvious from purely geometric considerations. Since the β strands form an open β sheet, there are no geometric restrictions on the number of strands involved. In fact, the number varies from four to ten. Furthermore, the two β strands joined by a crossover connection need not be adjacent in the β sheet, although the β-α-β motif where the two β strands are adjacent is a preferred structural building block. In addition, there can be mixed β sheets in which hairpin connections give rise to some antiparallel β strands mixed with the parallel β strands. All these variations occur in actual structures, some of which are illustrated in Figure 4.14a–d. There are thus many variations on the regular arrangement of six parallel β strands (see Figure 4.1b).

The positions of active sites can be predicted in α/β structures

We have described a general relationship between structure and function for the α/β-barrel structures. They all have the active site at the same position with respect to their common structure in spite of having different functions as well as different amino acid sequences. We can now ask if similar relationships also occur for the open α/β-sheet structures in spite of their much greater variation in structure. Can the position of the active sites be predicted from the structures of many open-sheet α/β proteins?

In almost every one of the more than 100 different known α/β structures of this class the active site is at the carboxy edge of the β sheet. Functional residues are provided by the loop regions that connect the carboxy end of the β strands with the amino end of the α helices. In this one respect a fundamental similarity therefore exists between the α/β-barrel structures and the open α/β-sheet structures.

The general shapes of the active sites are quite different, however. Open α/β structures cannot form funnel-shaped active sites like the barrel structures. Instead, they form crevices at the edge of the β sheet. Such crevices occur when there are two adjacent connections that are on opposite sides of the β sheet. One of the loop regions in these two connections goes out from

Petitioner Merck, Ex. 1014, p. 57

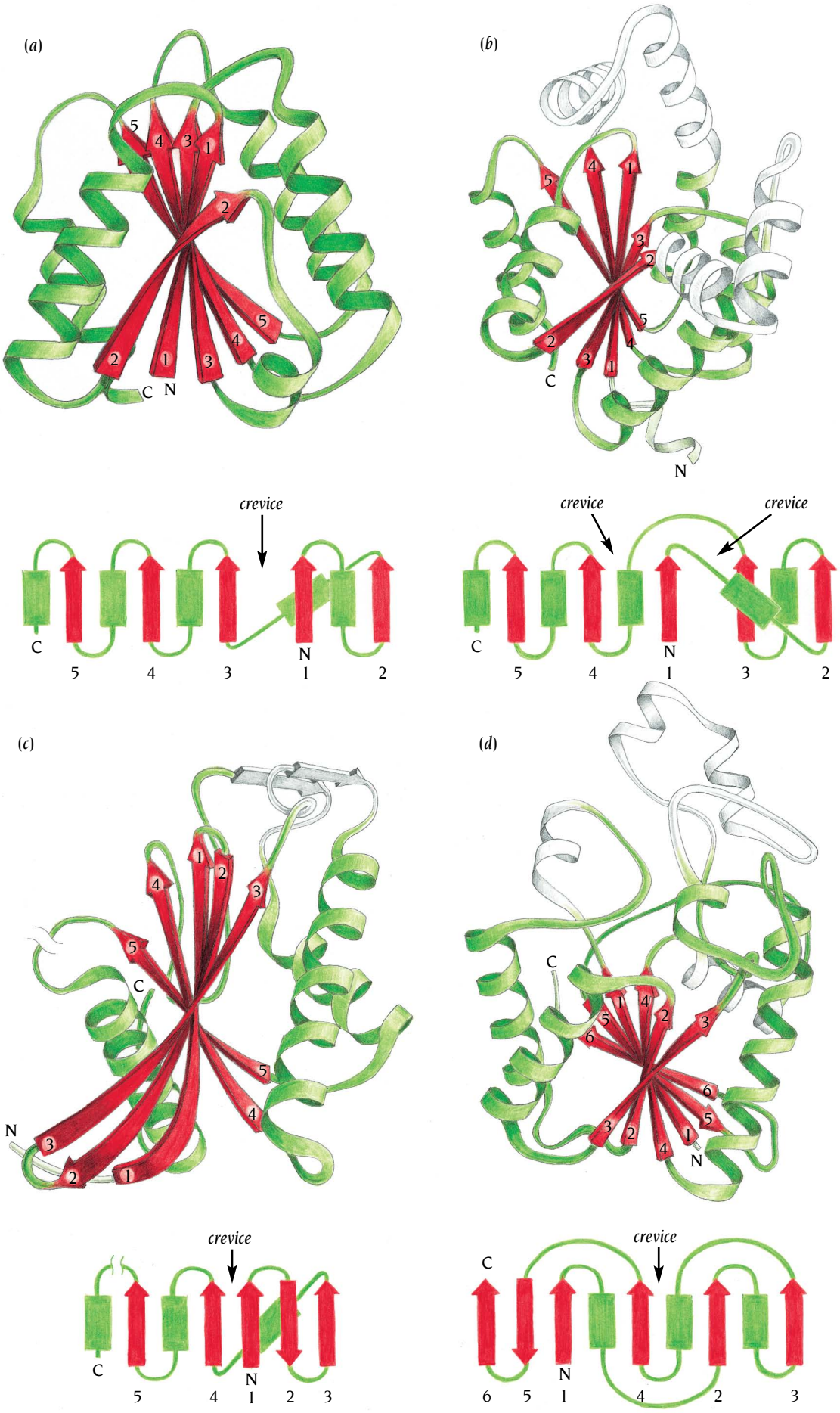


Figure 4.14 Examples of different types of open twisted α/β structures. Both schematic and topological diagrams are given. In the topological diagrams, arrows denote strands of β sheet and rectangles denote α helices. (a) The FMN-binding redox protein flavodoxin. (b) The enzyme adenylate kinase, which catalyzes the reaction AMP + ATP ⇌ 2 ADP. The structure was determined to 3.0 Å resolution in the laboratory of Georg Schulz in Heidelberg, Germany. (c) The ATP-binding domain of the glycolytic enzyme hexokinase, which catalyzes the phosphorylation of glucose. The structure was determined to 2.8 Å resolution in the laboratory of Tom Steitz, Yale University. (d) The glycolytic enzyme phosphoglycerate mutase, which catalyzes transfer of a phosphoryl group from carbon 3 to carbon 2 in phosphoglycerate. The structure was determined to 2.5 Å resolution in the laboratory of Herman Watson, Bristol University, UK. (Adapted from J. Richardson.)

Petitioner Merck, Ex. 1014, p. 58

its β strand above the β sheet and the other below, creating a crevice between them (see Figure 4.13). The active site or part of it is usually found in such a crevice. The position of such crevices is determined by the topology of the β sheet and can be predicted from a topology diagram. The crevices occur when the strand order is reversed, and can be easily identified in a topology diagram as the place where connections from the carboxy ends of two adjacent β strands go in opposite directions, one to the left and one to the right. Let us examine the first two diagrams given in Figure 4.14.

The first structure, flavodoxin (Figure 4.14a), has one such position, between strands 1 and 3. The connection from strand 1 goes to the right and that from strand 3 to the left. In the schematic diagram in Figure 4.14a we can see that the corresponding α helices are on opposite sides of the β sheet. The loops from these two β strands, 1 and 3, to their respective α helices form the major part of the binding cleft for the coenzyme FMN (flavin mononucleotide).

The second structure, adenylate kinase (Figure 4.14b), has two such positions, one on each side of β strand 1. The connection from strand 1 to strand 2 goes to the right, whereas the connection from the flanking strands 3 and 4 both go to the left. Crevices are formed between β strands 1 and 3 and between strands 1 and 4. One of these crevices forms part of an AMP-binding site, and the other crevice forms part of an ATP-binding site that catalyzes the formation of ADP from AMP and ATP.

Such positions in a topology diagram are called topological switch points. It was postulated in 1980 by Carl Branden, in Uppsala, Sweden, that the position of active sites could be predicted from such switch points. Since then at least one part of the active site has been found in crevices defined by such switch points in almost all new α/β structures that have been determined. Thus we can predict the approximate position of the active site and possible loop regions that form this site in α/β proteins. This is in contrast to proteins of the other two main classes—α-helical proteins and antiparallel β proteins—where no such predictive rules have been found. We will now examine a few examples that illustrate the relationship between the topology diagrams of some α/β proteins, their switch points, and the active-site residues. These examples have been chosen because they represent different types of α/β open-sheet structures.

Tyrosyl-tRNA synthetase has two different domains (α/β + α)

One of the crucial steps in protein synthesis is performed by the group of enzymes called aminoacyl-tRNA synthetases. These enzymes connect each amino acid with its specific transfer RNA molecule in a two-step reaction. First the amino acid is activated by ATP to give an enzyme-bound amino acid adenylate; then this complex is attacked by the tRNA to give the aminoacyl-tRNA.

The structure of the synthetase specific for the amino acid tyrosine was determined to 2.7 Å resolution in the laboratory of David Blow in London. Figure 4.15 shows a schematic diagram of the first 320 residues of a single subunit of this dimeric molecule. The last 100 residues are disordered in the crystal and are not visible. There are essentially two different domains, one

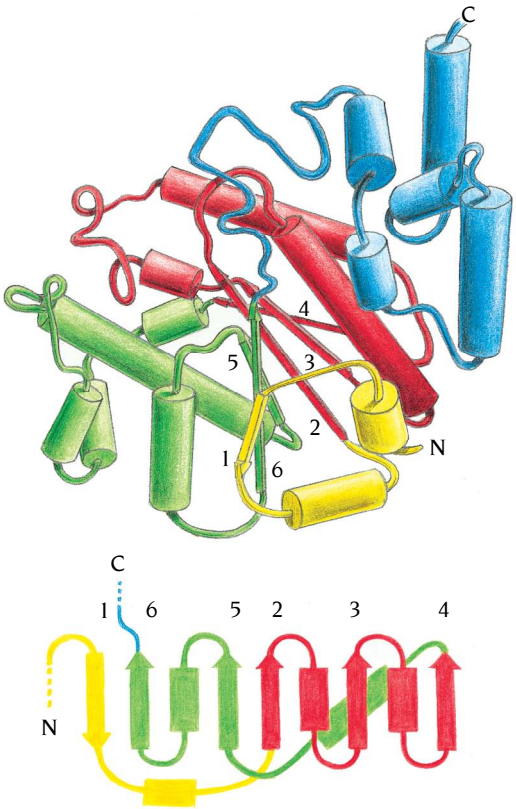


Figure 4.15 Schematic diagram of the enzyme tyrosyl-tRNA synthetase, which couples tyrosine to its cognate transfer RNA. The central region of the catalytic domain (red and green) is an open twisted α/β structure with five parallel β strands. The active site is formed by the loops from the carboxy ends of β strands 2 and 5. These two adjacent strands are connected to α helices on opposite sides of the β sheet. Where more than one α helix connects two β strands (for example, between strands 4 and 5), they are represented as one rectangle in the topology diagram. (Adapted from T.N. Bhat et al., *J. Mol. Biol.* 158: 699–709, 1982.)

Petitioner Merck, Ex. 1014, p. 59

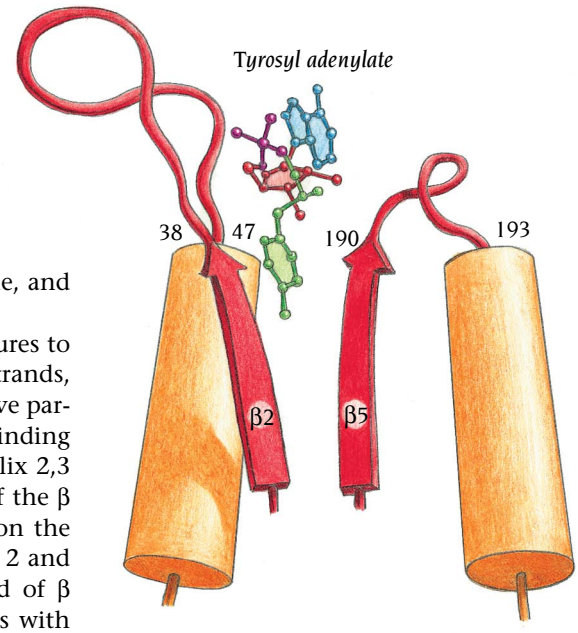


Figure 4.16 A schematic view of the active site of tyrosyl-tRNA synthetase. Tyrosyl adenylate, the product of the first reaction catalyzed by the enzyme, is bound to two loop regions: residues 38–47, which form the loop after β strand 2, and residues 190–193, which form the loop after β strand 5. The tyrosine and adenylate moieties are bound on opposite sides of the β sheet outside the carboxy ends of β strands 2 and 5.

α/β domain (red and green in Figure 4.15) that binds ATP and tyrosine, and one α-helical domain (blue), the function of which is not known.

Let us now apply the rules for predicting active sites of α/β structures to the topological diagram shown in Figure 4.15. The β sheet has six strands, one of which, number 1, is antiparallel to the others. The remaining five parallel β strands are arranged in a way rather similar to the nucleotide-binding fold (see Figure 4.1b), but here the strand order is 6 5 2 3 4. Alpha helix 2,3 (which connects β strands 2 and 3) and α helix 3,4 are on one side of the β sheet (red helices in Figure 4.15), whereas α helices 4,5 and 5,6 are on the other side (green helices). The switch point is thus between β strands 2 and 5. We would predict that the active site is outside the carboxy end of β strands 2 and 5 and that the loop regions that connect these strands with their respective α helices participate in binding the substrates. These loop regions comprise residues 38–47 and 190–193, respectively. The active site has been identified in the crystal structure by diffusing tyrosine and ATP into the crystals. The enzyme molecules in the crystals are active, so tyrosyl adenylate is formed, but because no tRNA is present, it stays bound to the enzyme.

The position of this bound tyrosyl adenylate was determined from an electron density map of the complex just after the predictive rules were formulated. The region where it binds proved to be as predicted. Loop regions 38–47, after β strand 2, and 190–193, after β strand 5, line a cleft where the substrate binds (Figure 4.16). The phosphate and the sugar moieties are hydrogen-bonded to the main-chain nitrogen atoms of residues 38 and 192, respectively. This part of the substrate is thus very close to the switch point. The substrate straddles the edge of the β sheet so that the tyrosine and adenine ends are on opposite sides of the β sheet. The substrate also interacts with some of the other regions at this end of the β sheet, especially residues 173–177 of the α helix that connects β strands 4 and 5 and, in addition, some residues within β strand 2. Figure 4.17 shows a schematic diagram of the position of bound tyrosine (red) in relation to these regions of the protein, and Figure 4.18 gives the important hydrogen bonds to the substrate, knowledge of which formed the basis for the beautiful site-directed mutagenesis experiments on this system by Alan Fersht in London.

Carboxypeptidase is an α/β protein with a mixed β sheet

Carboxypeptidases are zinc-containing enzymes that catalyze the hydrolysis of polypeptides at the C-terminal peptide bond. The bovine enzyme form A is a monomeric protein comprising 307 amino acid residues. The structure was determined in the laboratory of William Lipscomb, Harvard University, in 1970 and later refined to 1.5 Å resolution. Biochemical and x-ray studies have shown that the zinc atom is essential for catalysis by binding to the carbonyl oxygen of the substrate. This binding weakens the C′=O bond by

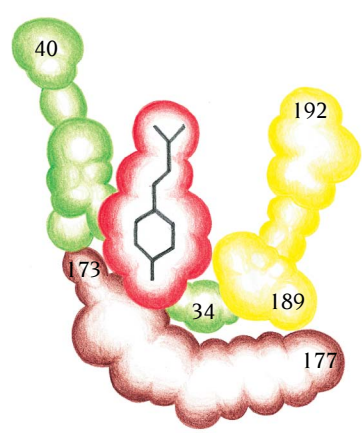


Figure 4.17 Schematic diagram of bound tyrosine to tyrosyl-tRNA synthetase. Colored regions correspond to van der Waals radii of atoms within a layer of the structure through the tyrosine ring. Red is bound tyrosine; green is the end of β strand 2 and the beginning of the following loop region; yellow is the loop region 189–192; and brown is part of the α helix in loop region 173–177.

Petitioner Merck, Ex. 1014, p. 60

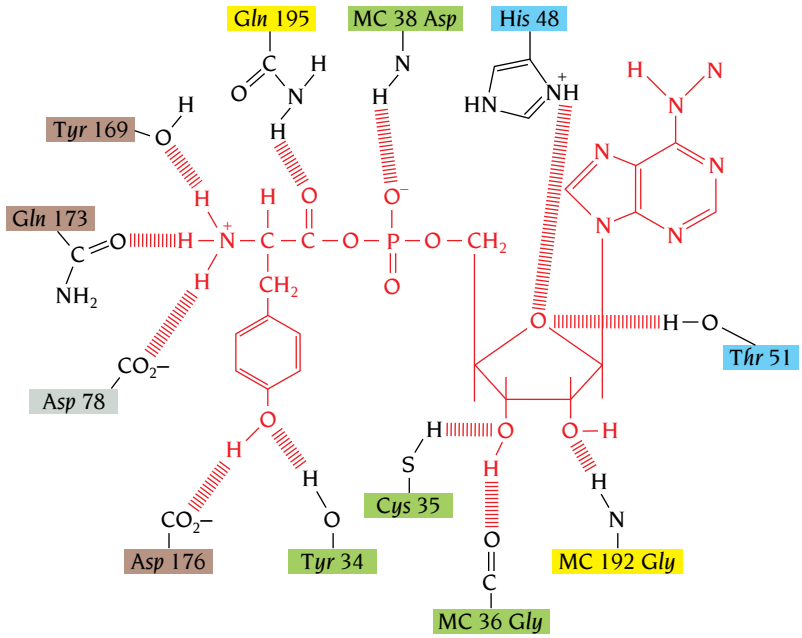


Figure 4.18 Side chains of the tyrosyl-tRNA synthetase that form hydrogen bonds to tyrosyl adenylate. Green residues are from β strand 2 and the following loop regions, yellow residues are from the loop after β strand 5, and brown residues are from the α helix before β strand 5. (Adapted from T. Wells and A. Fersht, *Nature* 316: 656–657, 1985.)

abstracting electrons from the carbon atom and thus facilitates cleavage of the adjacent peptide bond. Carboxypeptidase is a large single domain structure comprising a mixed β sheet of eight β strands (Figure 4.19) with α helices on both sides. Some of the loop regions are very long and curl around the central theme of the structure.

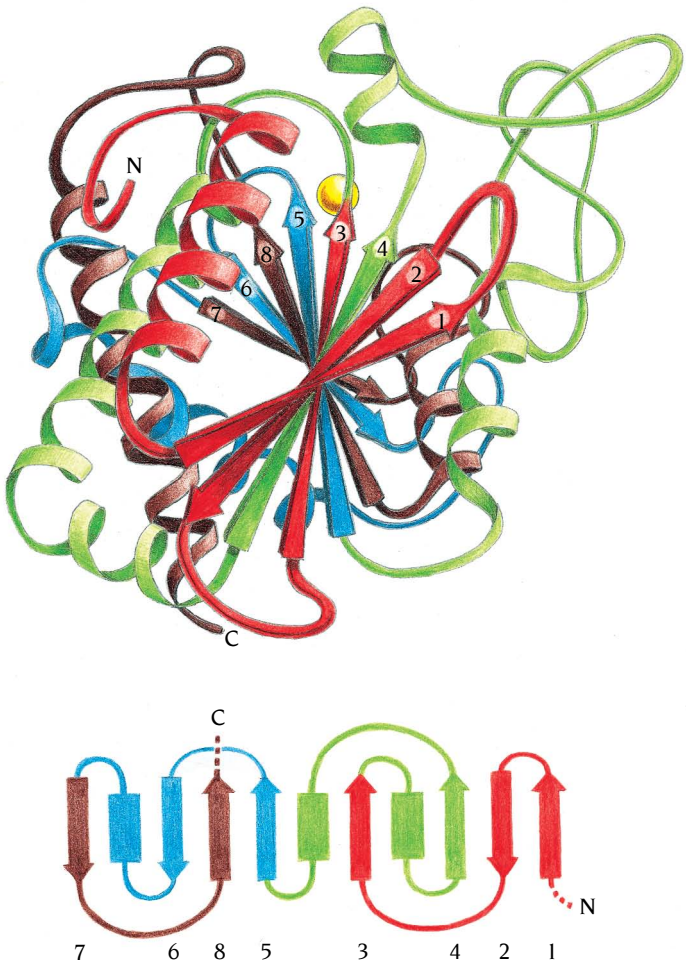


Figure 4.19 Schematic and topological diagrams for the structure of the enzyme carboxypeptidase. The central region of the mixed β sheet contains four adjacent parallel β strands (numbers 8, 5, 3, and 4), where the strand order is reversed between strands 5 and 3. The active-site zinc atom (yellow circle) is bound to side chains in the loop regions outside the carboxy ends of these two β strands. The first part of the polypeptide chain is red, followed by green, blue, and brown. (Adapted from J. Richardson.)

Petitioner Merck, Ex. 1014, p. 61

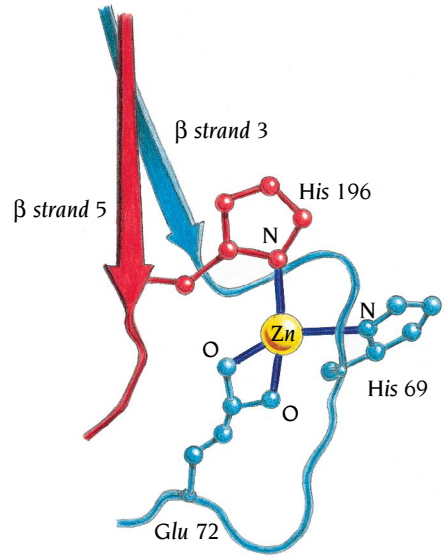


Figure 4.20 Detailed view of the zinc environment in carboxypeptidase. The active-site zinc atom is bound to His 69 and Glu 72, which are part of the loop region outside β strand 2. In addition, His 196, which is the last residue of β strand 5, also binds the zinc.

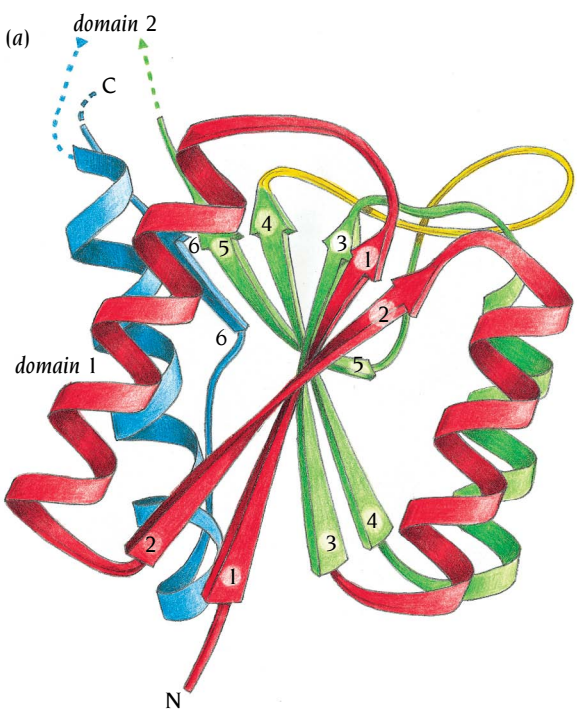
The four central strands of the β sheet are parallel and have the strand order 8 5 3 4 (see Figure 4.19). The strand order is thus reversed once, and there is a switch point in the middle of this β sheet between β strands 5 and 3 where we would expect the active site to be located.

This is precisely where the catalytically essential zinc atom is found. This zinc atom is located precisely at this switch point, where it is firmly anchored to the protein by three side-chain ligands, His 69, Glu 72, and His 196 (Figure 4.20). The last residue of β strand 3 is residue 66, so the two zinc ligands His 69 and Glu 72 are at the beginning of the loop region that connects this β strand with its corresponding α helix. The last residue of β strand 5 is the third zinc ligand, His 196.

In this structure the loop regions adjacent to the switch point do not provide a binding crevice for the substrate but instead accommodate the active-site zinc atom. The essential point here is that this zinc atom and the active site are in the predicted position outside the switch point for the four central parallel β strands, even though these β strands are only a small part of the total structure. This sort of arrangement, in which an active site formed from parallel β strands is flanked by antiparallel β strands, has been found in a number of other α/β proteins with mixed β sheets.

Arabinose-binding protein has two similar α/β domains

The arabinose-binding protein is one of the group of proteins that occur in the periplasmic space between the inner and outer cell membranes of Gram-negative bacteria such as *E. coli.* These proteins are components of active transport systems for various sugars, amino acids, and ions. Arabinose-binding protein is involved in arabinose transport. It is a single polypeptide chain of 306 amino acids folded into two domains of similar structure and topology (Figure 4.21), as was ascertained in the laboratory of Florante Quiocho in Houston, Texas, by a structure determination of the protein with bound



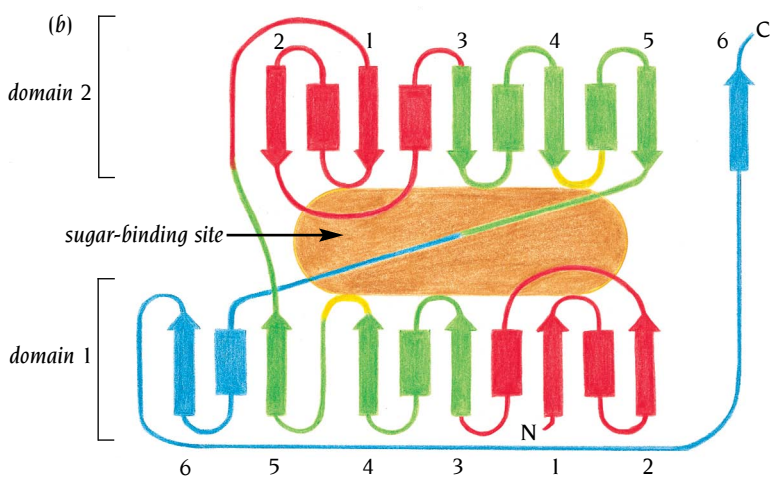


Figure 4.21 The polypeptide chain of the arabinose-binding protein in *E. coli* contains two open twisted α/β domains of similar structure. A schematic diagram of one of these domains is shown in (a). The two domains are oriented such that the carboxy ends of the parallel β strands face each other on opposite sides of a crevice in which the sugar molecule binds, as illustrated in the topology diagram (b). [(a) Adapted from J. Richardson.]

Petitioner Merck, Ex. 1014, p. 62

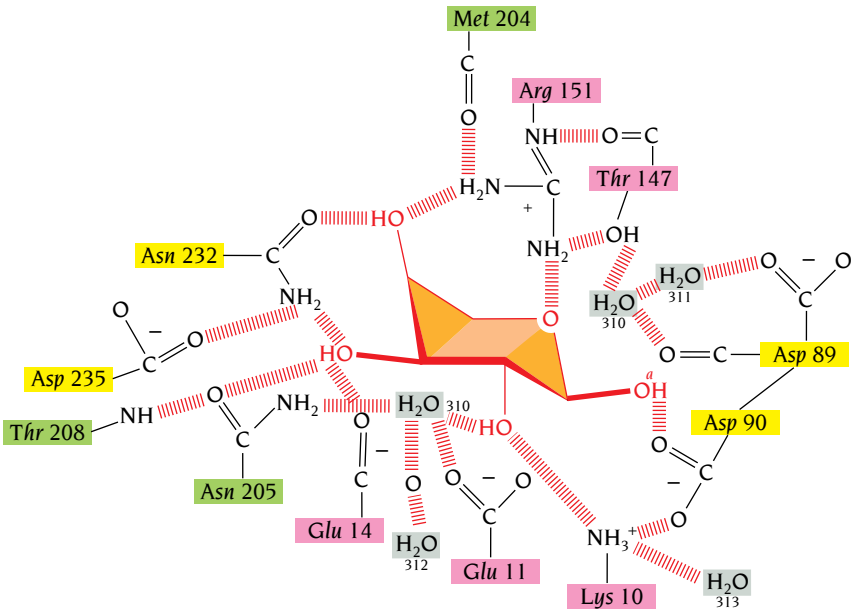


Figure 4.22 Schematic diagram of the complex networks of hydrogen bonds formed by polar side chains from the arabinose-binding protein and L-arabinose. The residues that interact with the sugar are in turn hydrogen-bonded to each other or to other residues or isolated water molecules. The pink and green residues are in loop regions that, from the topology diagram, are predicted to form the binding site. The yellow residues are from adjacent loop regions. (Adapted from C.F. Sams et al., *Nature* 310: 429–430, 1984.)

arabinose at 1.7 Å resolution. The first five β strands in both domains are parallel; the strand order is reversed once, and the switch points are between β strands 1 and 3 in both domains.

A schematic diagram relating binding to the topology of the domains is shown in Figure 4.21b. Twelve amino acid residues from both domains are involved in forming a complicated network of hydrogen bonds to the oxygen atoms of the bound arabinose (Figure 4.22). We predict from the topology of these domains that loop residues from β strands 1 and 3 of both domains should participate in these interactions. Five residues in loop regions after β strand 1 in both domains (pink residues in Figure 4.22) and three residues after β strand 3 in the second domain (green residues in Figure 4.22) participate in binding. The remaining four residues (yellow residues in Figure 4.22) are all from loop regions after β strand 4 in both domains. These are one β strand removed from the switch point. Both domains thus participate approximately equally in binding the sugar at the carboxy edge of the β sheets, outside the switch point, where the strand order is reversed.

A number of proteins consist of two α/β open-sheet domains formed from a single polypeptide chain, as in arabinose-binding protein. In almost all these cases the active sites are found in cleft regions between these two domains. The domains are oriented in such a way that the carboxy edge of both β sheets points toward the active site. Loop regions adjacent to the switch points of both domains participate in forming the active site. In enzymatic reactions where two different substrates participate, they are bound to different domains and are brought together for catalytic reactions by the orientation of these domains. In other proteins the two domains bind different regions of the same ligand. Sugar-binding proteins from bacteria are examples of this second case.

Conclusion

Alpha/beta (α/β) structures are the most frequent and most regular of the protein structures. They fall into three classes: the first class comprises a central core of usually eight parallel β strands arranged close together like the staves of a barrel, surrounded by α helices; the second class comprises an open twisted parallel or mixed β sheet with α helices on both sides of the β sheet; and the third class is formed by leucine-rich motifs in which a large number of parallel β strands form a curved β sheet with all the α helices on the outside of this sheet.

Petitioner Merck, Ex. 1014, p. 63

The α/β-barrel structure is one of the largest and most regular of all domain structures, comprising about 250 amino acids. It has so far been found in more than 20 different proteins, with completely different amino acid sequences and different functions. They are all enzymes that are modeled on this common scaffold of eight parallel β strands surrounded by eight α helices. They all have their active sites in very similar positions, at the bottom of a funnel-shaped pocket created by the loops that connect the carboxy end of the β strands with the amino end of the α helices. The specific enzymatic activity is, in each case, determined by the lengths and amino acid sequences of these loop regions, which do not contribute to the stability of the fold.

The horseshoe structure is formed by homologous repeats of leucine-rich motifs, each of which forms a β-loop-α unit. The units are linked together such that the β strands form an open curved β sheet, like a horseshoe, with the α helices on the outside of the β sheet and the inside exposed to solvent. The invariant leucine residues of these motifs form the major part of the hydrophobic region between the α helices and the β sheet.

The open α/β-sheet structures vary considerably in size, number of β strands, and strand order. Independent of these variations, they all have their active sites at the carboxy edge of the β strands, and these active sites are lined by the loop regions that connect the β strands with the α helices. In this respect, they are similar to the α/β-barrel structures. However, the active-site regions are created differently in open structures. They are formed in those regions outside the carboxy edge of the β sheet, where two adjacent loops are on opposite sides of the β sheet. The positions of these regions can be predicted from topology diagrams. The rules that relate the general position of functional binding sites to the overall structure of the protein are thus known for β barrels and open-sheet α/β proteins.

Selected readings

General

Babbit, P.C., et al. A functionally diverse enzyme superfamily that abstracts the α protons of carboxylic acids. *Science* 267: 1159–1161, 1995.

Brändén, C.-I. Relation between structure and function of α/β proteins. *Q. Rev. Biophys.* 13: 317–338, 1980.

Brenner, S.E. Gene duplications in *H. influenzae*. *Nature* 378: 140, 1995.

Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins. *J. Mol. Biol.* 156: 821–862, 1982.

Farber, G., Petsko, G.A. The evolution of α/β-barrel enzymes. *Trends Biochem. Sci.* 15: 228–234, 1990.

Janin, J., Chothia, C. Packing of α-helices onto β-pleated sheets and the anatomy of α/β-proteins. *J. Mol. Biol.* 143: 95–128, 1980.

Kajava, A.V., Vassart, G., Wodak, S.J. Modelling of the three-dimensional structure of proteins with the typical leucine-rich repeats. *Structure* 3: 867–877, 1995.

Lasters, I., et al. Structural principles of parallel β barrels in proteins. *Proc. Natl. Acad. Sci. USA* 85: 3338–3342, 1988.

Lasters, I., Wodak, S.J., Pio, F. The design of idealized α/β-barrels: analysis of β-sheet closure requirements. *Proteins* 7: 249–256, 1990.

Lesk, A.M., Branden, C.-I., Chothia, C. Structural principles of α/β barrel proteins: the packing of the interior of the sheet. *Proteins* 5: 139–148, 1989.

Murzin, A.G., Lesk, A.M., Chothia, C. Principles determining the structure of β sheet barrels in proteins. *J. Mol. Biol.* 236: 1369–1400, 1994.

Ohlsson, I., Nordström, B., Brändén, C.-I. Structural and functional similarities within the coenzyme binding domains of dehydrogenases. *J. Mol. Biol.* 89: 339–354, 1974.

Richardson, J.S. β-sheet topology and the relatedness of proteins. *Nature* 268: 495–500, 1977.

Richardson, J.S. Handedness of crossover connections in β sheets. *Proc. Natl. Acad. Sci. USA* 73: 2619–2623, 1976.

Sternberg, M.J.E., et al. Analysis and prediction of structural motifs in the glycolytic enzymes. *Phil. Trans. R. Soc. Lond.* B293: 177–189, 1981.

Sternberg, M.J.E., Thornton, J.M. On the conformation of proteins: the handedness of the connection between parallel β-strands. *J. Mol. Biol.* 110: 269–283, 1977.

Specific structures

Brick, P., Bhat, T.N., Blow, D.M. Structure of tyrosyl-tRNA synthetase refined at 2.3 Å resolution. Interaction of the enzyme with the tyrosyl adenylate intermediate. *J. Mol. Biol.* 208: 83–98, 1988.

Petitioner Merck, Ex. 1014, p. 64

Campbell, J.W., Watson, H.C., Hodgson, G.I. Structure of yeast phosphoglycerate mutase. *Nature* 250: 301–303, 1974.

Carrel, H.L., et al. X-ray structure of D-xylose isomerase from *Streptomyces rubiginosus* at 4 Å resolution. *J. Biol. Chem.* 259: 3230–3236, 1984.

Dreusicke, D., Karplus, P.A., Schulz, G.E. Refined structure of porcine adenylate kinase at 2.1 Å resolution. *J. Mol. Biol.* 199: 359–371, 1988.

Eklund, H., et al. Three-dimensional structure of horse liver alcohol dehydrogenase at 2.4 Å resolution. *J. Mol. Biol.* 102: 27–59, 1976.

Gilliland, G.L., Quiocho, F.A. Structure of the L-arabinose-binding protein from *Escherichia coli* at 2.4 Å resolution. *J. Mol. Biol.* 146: 341–362, 1981.

Goldman, A., Ollis, D.L., Steitz, T.A. Crystal structure of muconate lactonizing enzyme at 3 Å resolution. *J. Mol. Biol.* 194: 143–153, 1987.

Hofmann, B.E., Bender, H., Schulz, G.E. Three-dimensional structure of cyclodextrin glycosyltransferase from *Bacillus circulans* at 3.4 Å resolution. *J. Mol. Biol.* 209: 793–800, 1989.

Hyde, C.C., et al. Three-dimensional structure of the tryptophan synthase α_2 β_2 multienzyme complex from *Salmonella typhimurium*. *J. Biol. Chem.* 263: 17857–17871, 1988.

Knight, S., Andersson, I., Brändén, C.-I. Crystallographic analysis of ribulose-1,5-bisphosphate carboxylase from spinach at 2.4 Å resolution. Subunit interactions and active site. *J. Mol. Biol.* 215: 113–160, 1990.

Kobe, B., Deisenhofer, J. Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats. *Nature* 366: 751–756, 1993.

Lebioda, L., Stec, B., Brewer, J.M. The structure of yeast enolase at 2.5 Å resolution. An 8-fold $\beta + \alpha$ barrel with a novel $\beta\beta\alpha\alpha(\beta\alpha)_6$ topology. *J. Biol. Chem.* 264: 3685–3693, 1989.

Lim, L.W., et al. Three-dimensional structure of the iron-sulfur flavoprotein trimethylamine dehydrogenase at 2.4 Å resolution. *J. Biol. Chem.* 261: 15140–15146, 1986.

Lindqvist, Y. Refined structure of spinach glycolate oxidase at 2 Å resolution. *J. Mol. Biol.* 209: 151–166, 1989.

Mancia, F., et al. How coenzyme B_{12} radicals are generated: the crystal structure of methylmalonyl–coenzyme A mutase at 2 Å resolution. *Structure* 4: 339–350, 1996.

Matsuura, Y., et al. Structure and possible catalytic residues of taka-amylase A. *J. Biochem.* 95: 697–702, 1984.

Mavridis, I.M., et al. Structure of 2-keto-3-deoxy-6-phosphogluconate aldolase at 2.8 Å resolution. *J. Mol. Biol.* 162: 419–444, 1982.

Muirhead, H., et al. The structure of cat muscle pyruvate kinase. *EMBO J.* 5: 475–481, 1986.

Neidhart, D.J., et al. Mandelate racemase and muconate lactonizing enzyme are mechanistically distinct and structurally homologous. *Nature* 347: 692–694, 1990.

Priestle, J.P., et al. Three-dimensional structure of the bifunctional enzyme N-(5′-phosphoribosyl) anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 84: 5690–5694, 1987.

Rees, D.C., Lewis, M., Lipscomb, W.N. Refined crystal structure of carboxypeptidase A at 1.54 Å resolution. *J. Mol. Biol.* 168: 367–387, 1983.

Rouvinen, J., et al. Three-dimensional structure of cellobiohydrolase II from *Trichoderma resei*. *Science* 249: 380–386, 1990.

Schneider, G., Lindqvist, Y., Lundqvist, T. Crystallographic refinement and structure of ribulose-1,5-bisphosphate carboxylase from *Rhodospirillum rubrum* at 1.7 Å resolution. *J. Mol. Biol.* 211: 989–1008, 1990.

Steitz, T.A., et al. High resolution x-ray structure of yeast hexokinase, an allosteric protein exhibiting a non-symmetric arrangement of subunits. *J. Mol. Biol.* 104: 197–222, 1976.

Sygusch, J., Beaudry, D., Allaire, M. Molecular architecture of rabbit skeletal muscle aldolase at 2.7 Å resolution. *Proc. Natl. Acad. Sci. USA* 84: 7846–7850, 1987.

Uhlin, U., Eklund, H. Structure of ribonucleotide reductase protein R1. *Nature* 370: 553–559, 1994.

Xia, Z.-X., et al. Three-dimensional structure of flavocytochrome b_2 from baker's yeast at 3.0 Å resolution. *Proc. Natl. Acad. Sci. USA* 84: 2629–2633, 1987.

Petitioner Merck, Ex. 1014, p. 65

Beta Structures

5

Antiparallel beta (β) structures comprise the second large group of protein domain structures. Functionally, this group is the most diverse; it includes enzymes, transport proteins, antibodies, cell surface proteins, and virus coat proteins. The cores of these domains are built up by β strands that can vary in number from four or five to over ten. The β strands are arranged in a predominantly antiparallel fashion and usually in such a way that they form two β sheets that are joined together and packed against each other.

The β sheets have the usual twist, and when two such twisted β sheets are packed together, they form a barrel-like structure (Figure 5.1). Antiparallel β structures, therefore, in general have a core of hydrophobic side chains inside the barrel provided by residues in the β strands. The surface is formed by residues from the loop regions and from the strands. The aim of this chapter is to examine a number of antiparallel β structures and demonstrate how these rather complex structures can be separated into smaller comprehensible motifs.

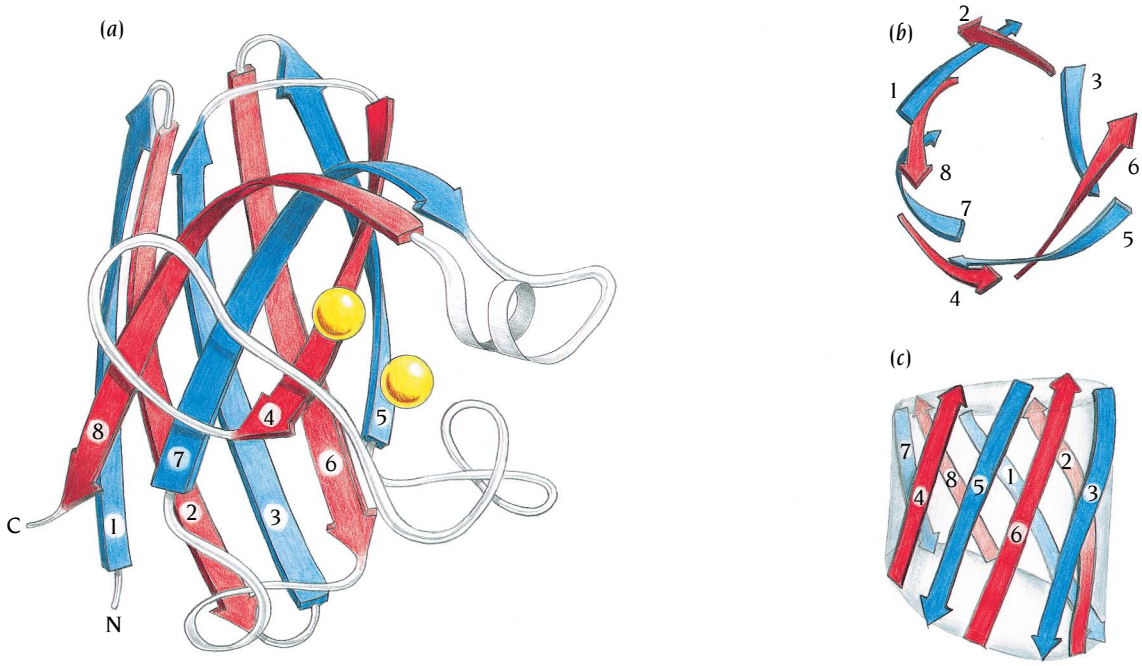


Figure 5.1 The enzyme superoxide dismutase (SOD). SOD is a β structure comprising eight antiparallel β strands (a). In addition, SOD has two metal atoms, Cu and Zn (yellow circles), that participate in the catalytic action: conversion of a superoxide radical to hydrogen peroxide and oxygen. The eight β strands are arranged around the surface of a barrel, which is viewed along the barrel axis in (b) and perpendicular to this axis in (c). [(a) Adapted from J.S. Richardson. The structure of SOD was determined in the laboratory of J.S. and D.R. Richardson, Duke University.]

Petitioner Merck, Ex. 1014, p. 67

In Chapter 2 we described the 24 different ways that two β-loop-β units can form a four-stranded β sheet. The number of possible ways to form antiparallel β-sheet structures rapidly increases as the number of strands increases. It is thus surprising, but reassuring, that the number of topologies actually observed is small and that most β structures fall into a few groups of common or similar topology. The three most frequently occurring groups—up-and-down barrels, Greek keys, and jelly roll barrels—can all be related to simple ways of connecting antiparallel β strands arranged in a barrel structure.

Up-and-down barrels have a simple topology

The simplest topology is obtained if each successive β strand is added adjacent to the previous strand until the last strand is joined by hydrogen bonds to the first strand and the barrel is closed (Figure 5.2). These are called **up-and-down β sheets** or **barrels**. The arrangement of β strands is similar to that in the α/β-barrel structures we have just described in Chapter 4, except that here the strands are antiparallel and all the connections are hairpins. The structural and functional versatility of even this simple arrangement will be illustrated by two examples.

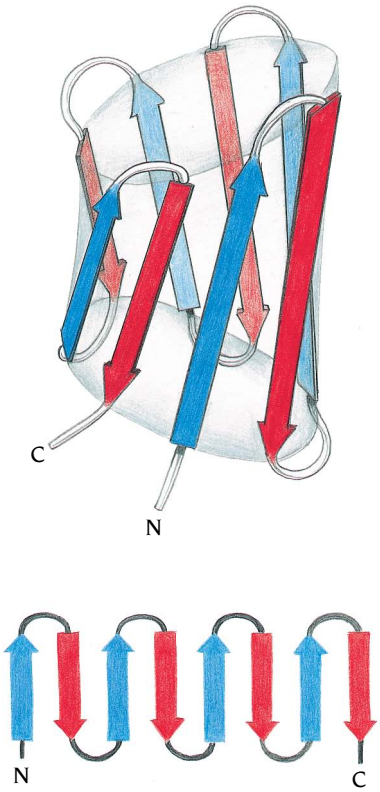


Figure 5.2 Schematic and topological diagrams of an up-and-down β barrel. The eight β strands are all antiparallel to each other and are connected by hairpin loops. Beta strands that are adjacent in the amino acid sequence are also adjacent in the three-dimensional structure of up-and-down barrels.

The retinol-binding protein binds retinol inside an up-and-down β barrel

The first example is the plasma-borne **retinol-binding protein**, RBP, which is a single polypeptide chain of 182 amino acid residues. This protein is responsible for transporting the lipid alcohol **vitamin A** (retinol) from its storage site in the liver to the various vitamin-A-dependent tissues. It is a disposable package in the sense that each RBP molecule transports only a single retinol molecule and is then degraded.

RBP is synthesized in the hepatocytes, where it picks up one molecule of retinol in the endoplasmic reticulum. Both its synthesis and its secretion from the hepatocytes to the plasma are regulated by retinol. In plasma, the

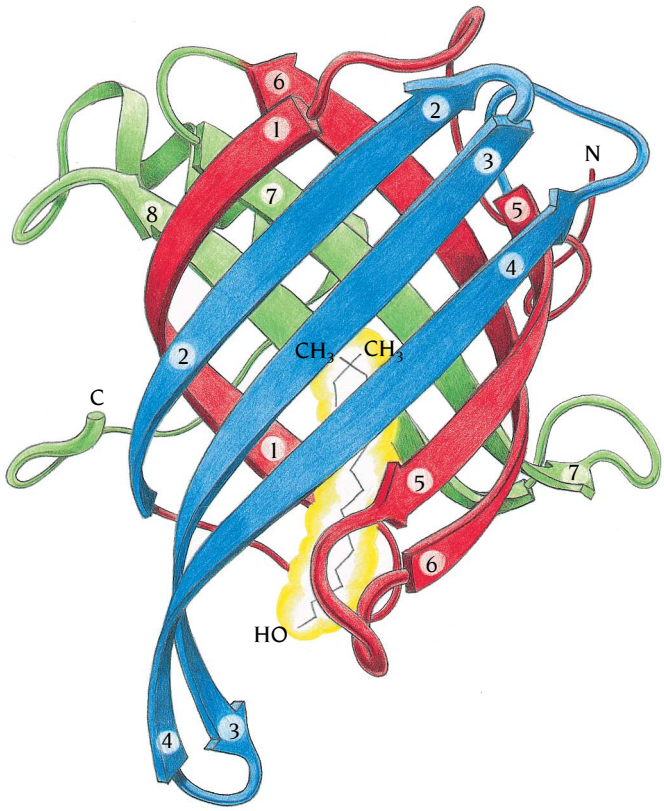


Figure 5.3 Schematic diagram of the structure of human plasma retinol-binding protein (RBP), which is an up-and-down β barrel. The eight antiparallel β strands twist and curl such that the structure can also be regarded as two β sheets (green and blue) packed against each other. Some of the twisted β strands (red) participate in both β sheets. A retinol molecule, vitamin A (yellow), is bound inside the barrel, between the two β sheets, such that its only hydrophilic part (an OH tail) is at the surface of the molecule. The topological diagram of this structure is the same as that in Figure 5.2. (Courtesy of Alwyn Jones, Uppsala, Sweden.)

Petitioner Merck, Ex. 1014, p. 68

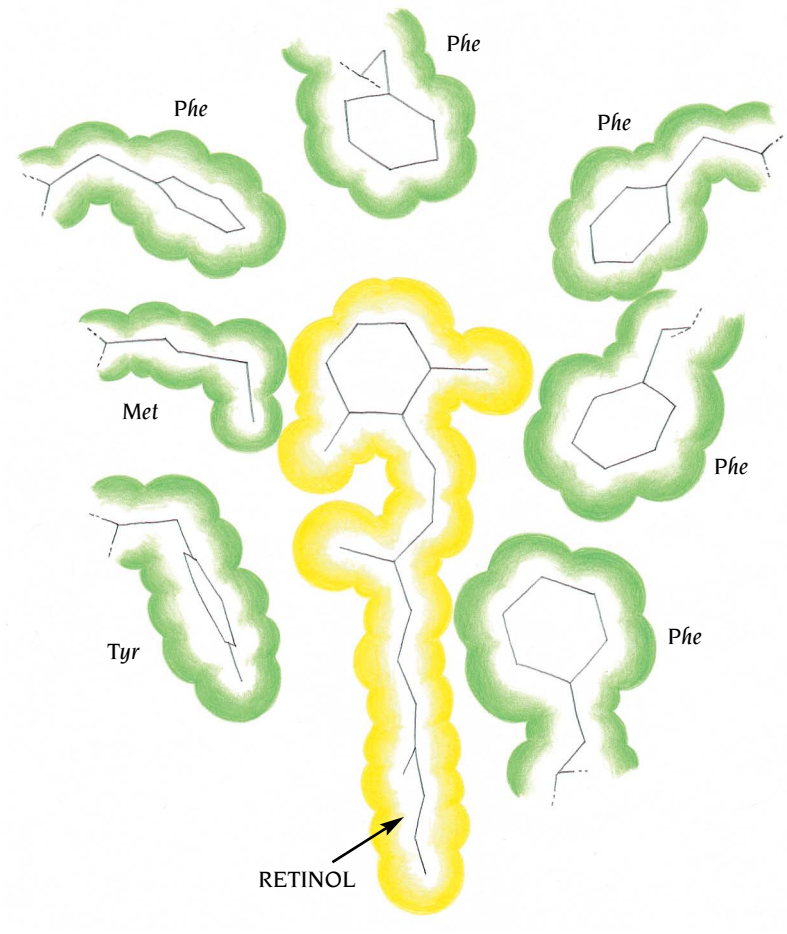


Figure 5.4 The binding site for retinol inside the RBP barrel is lined with hydrophobic residues. They provide a hydrophobic surrounding for the hydrophobic part of the retinol molecule.

RBP-retinol complex binds to a larger protein molecule, prealbumin, which further stabilizes it and prevents its loss via the kidney. Recognition of this complex by a cell-surface receptor causes RBP to release the retinol and, as a result, to undergo a conformational change that drastically reduces its affinity for prealbumin. The free RBP molecule is then excreted through the kidney glomerus, reabsorbed in the proximal tubule cells, and degraded.

The structure of RBP with bound retinol has been determined in the laboratory of Alwyn Jones in Uppsala, Sweden to 2.0 Å resolution. Its most striking feature is a β-barrel core consisting of eight up-and-down antiparallel β strands as shown in Figure 5.3. In addition, there is a four-turn α helix at the carboxy end of the polypeptide chain that is packed against the outside of the β barrel. The β strands are curved and twisted, and the barrel is wrapped around the retinol molecule. One end of the barrel is open to the solvent whereas the other end is closed by tight side-chain packing: the tail of the retinol molecule is at the open end of the barrel.

The hydrophobic retinol molecule is packed against hydrophobic side chains from the β strands in the barrel's core (Figure 5.4). The structure of the apo-form where the retinol molecule is removed is also known. Surprisingly, there is almost no change in the barrel structure, and a large hole is left inside the barrel.

Amino acid sequence reflects β structure

On a large part of the surface of RBP (the front face in Figure 5.3), side chains from residues in the β strands are exposed to the solvent. This is achieved by alternating hydrophobic with polar or charged hydrophilic residues in the

Petitioner Merck, Ex. 1014, p. 69

strand no.	residue no.	amino acid sequence
2	41–48	– Ile – Val – Ala – Glu – Phe – Ser – Val – Asp –
3	53–60	– Met – Ser – Ala – Thr – Ala – Lys – Gly – Arg –
4	71–78	– Ala – Asp – Met – Val – Gly – Thr – Phe – Thr –

Figure 5.5 Amino acid sequence of β strands 2 3 4 in human plasma retinol-binding protein. The sequences are listed in such a way that residues which point into the barrel are aligned. These hydrophobic residues are arrowed and colored green. The remaining residues are exposed to the solvent.

amino acid sequences of the β strands; in other words, side chains of the β strands form the hydrophobic core of the barrel as well as part of the hydrophilic outer surface. Strands 2 3 4 of RBP clearly illustrate this arrangement (Figure 5.5, where the core residues are colored green). This structure also very clearly illustrates that an antiparallel β barrel is built up from two β sheets that are packed against each other (see Figure 5.3). Beta strands 1 2 3 4 5 6 (blue and red color) form one sheet, and strands 1 8 7 6 5 (green and red color) form the second sheet. Strands 1 5 6 thus contribute to both sheets by having sharp corners where they can turn over from one sheet to the other.

The retinol-binding protein belongs to a superfamily of protein structures

RBP is one member of a **superfamily** of proteins with different functions, marginally homologous amino acid sequences, but similar three-dimensional structures. This superfamily also includes an insect protein that binds a blue pigment, biliverdin, and β-lactoglobulin, a protein that is abundant in milk. All have polypeptide chains of approximately the same lengths that are wrapped into very similar up-and-down eight-stranded antiparallel β barrels. They all tightly bind hydrophobic ligands inside this barrel.

There is a second family of small lipid-binding proteins, the **P2 family**, which include among others cellular retinol- and fatty acid-binding proteins as well as a protein, P2, from myelin in the peripheral nervous system. However, members of this second family have ten antiparallel β strands in their barrels compared with the eight strands found in the barrels of the RBP superfamily. Members of the P2 family show no amino acid sequence homology to members of the RBP superfamily. Nevertheless, their three-dimensional structures have similar architecture and topology, being up-and-down β barrels.

Neuraminidase folds into up-and-down β sheets

A second example of up-and-down β sheets is the protein neuraminidase from influenza virus. Here the packing of the sheets is different from that in RBP. They do not form a simple barrel but instead six small sheets, each with four β strands, which are arranged like the blades of a six-bladed propeller. Loop regions between the β strands form the active site in the middle of one side of the propeller. Other similar structures are known with different numbers of the same motif arranged like propellers with different numbers of blades such as the G-proteins discussed in Chapter 13.

Influenza virus is an RNA virus with an outer lipid envelope. There are two viral proteins anchored in this membrane, neuraminidase and hemagglutinin. They are both transmembrane proteins with a few residues inside the membrane and a transmembrane region followed by a stalk and a headpiece outside the membrane. The heads are exposed on the surface of the virion and thus provide the antigenic determinants of this epidemic virus. The function of hemagglutinin, which is glycosylated, is to mediate the binding of virus particles to host cells by recognizing and binding to sialic acid residues on glycoproteins of the cell membrane, as we shall discuss in more detail at the end of this chapter.

The role of the viral neuraminidase, conversely, seems to be to facilitate the release of progeny virions from infected cells by cleaving sialic acid

Petitioner Merck, Ex. 1014, p. 70

residues from the carbohydrate side chains both of the viral hemagglutinin and of the glycosylated cellular membrane proteins. This helps prevent progeny virions from binding to and reinfecting the cells from whose surface they have just budded. From the point of view of the viruses, reinfecting an already infected cell is, of course, a waste of time.

The neuraminidase molecule is a homotetramer made up of four identical polypeptide chains, each of around 470 amino acids; the exact number varies depending on the strain of the virus. If influenza virus is treated with the proteolytic enzyme pronase, the head of the neuraminidase, which is soluble, is cleaved off from the stalk projecting from the viral envelope. The soluble head, comprising four subunits of about 400 amino acids each, can be crystallized.

Folding motifs form a propeller-like structure in neuraminidase

The structure of these tetrameric neuraminidase heads was determined in the laboratory of Peter Colman in Parkville, Australia to 2.9 Å resolution. Each of the four subunits of the tetramer is folded into a single domain built up from six closely packed, similarly folded motifs. The motif is a simple up-and-down antiparallel β sheet of four strands (Figure 5.6). The strands have a rather large twist such that the directions of the first and the fourth strands differ by 90°. To a first very rough approximation the six motifs are arranged within each subunit with an approximate sixfold symmetry around an axis through the center of the subunit (Figure 5.7a). These six β sheets are arranged like six blades of a propeller.

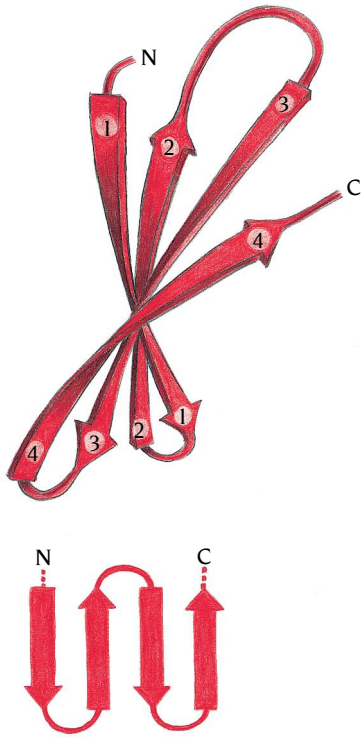


Figure 5.6 Schematic and topological diagrams of the folding motif in neuraminidase from influenza virus. The motif is built up from four antiparallel β strands joined by hairpin loops, an up-and-down open β sheet.

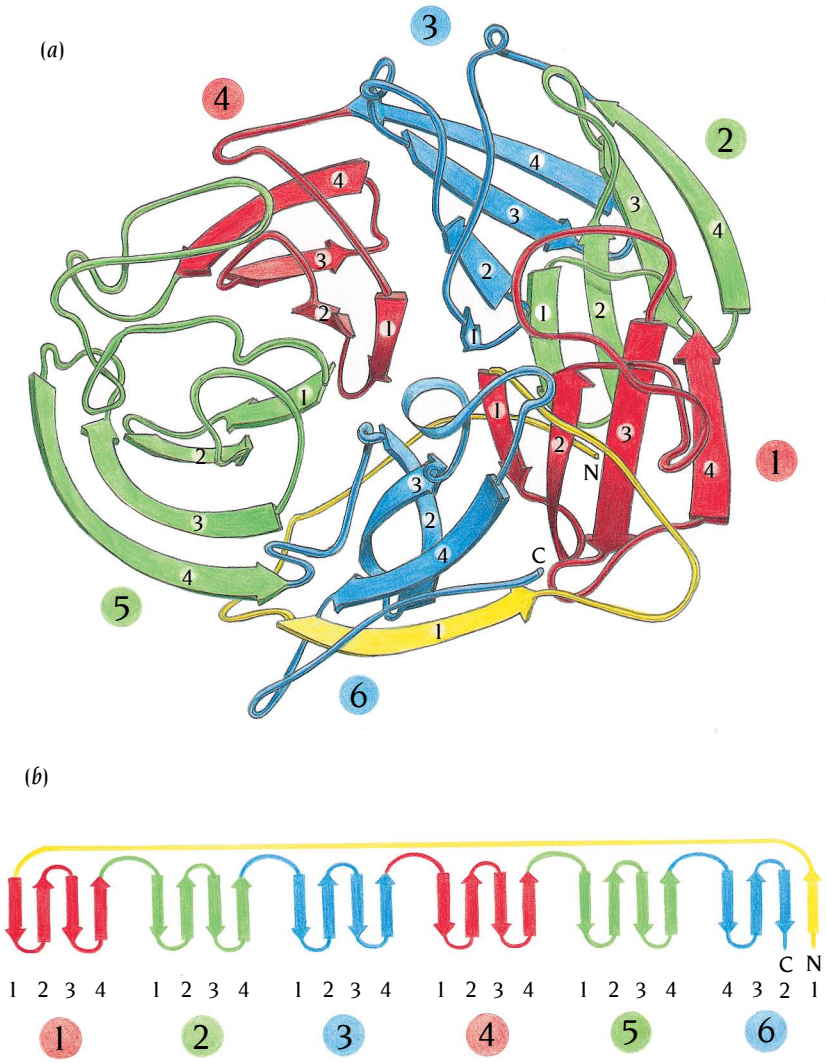


Figure 5.7 The subunit structure of the neuraminidase headpiece (residues 84–469) from influenza virus is built up from six similar, consecutive motifs of four up-and-down antiparallel β strands (Figure 5.6). Each such motif has been called a propeller blade and the whole subunit structure a six-blade propeller. The motifs are connected by loop regions from β strand 4 in one motif to β strand 1 in the next motif. The schematic diagram (a) is viewed down an approximate sixfold axis that relates the centers of the motifs. Four such six-blade propeller subunits are present in each complete neuraminidase molecule (see Figure 5.8). In the topological diagram (b) the yellow loop that connects the N-terminal β strand to the first β strand of motif 1 is not to scale. In the folded structure it is about the same length as the other loops that connect the motifs. (Adapted from J. Varghese et al., *Nature* 303: 35–40, 1983.)

Petitioner Merck, Ex. 1014, p. 71

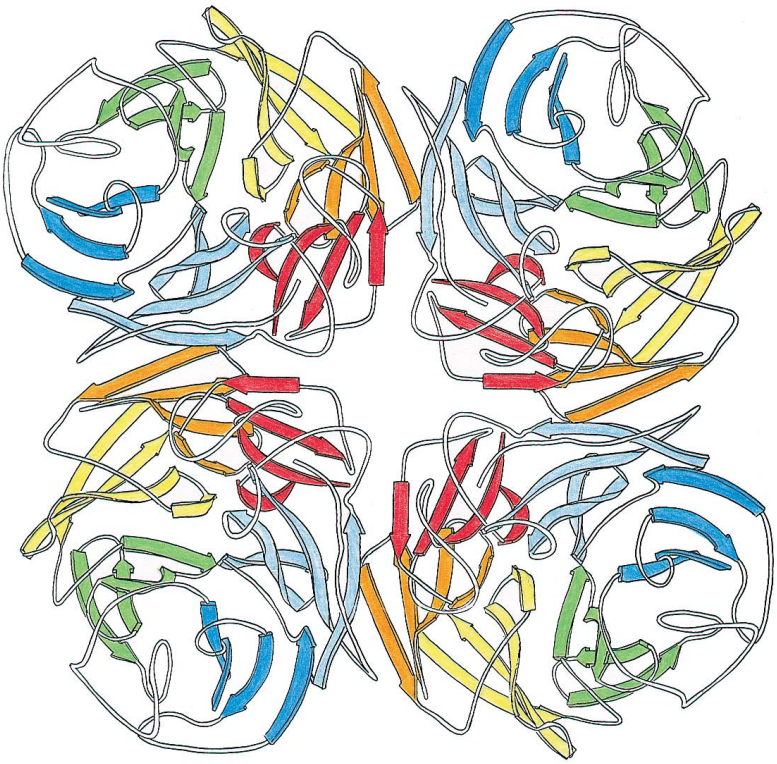


Figure 5.8 Schematic view down the fourfold axis of the tetrameric molecule of neuraminidase as it appeared on the cover of *Nature*, May 5, 1983.

In summary, the whole molecule has almost 1600 amino acid residues. It is composed of four identical polypeptide chains, each of which is folded into a superbarrel with 24 β strands (Figure 5.8). These 24 β strands are arranged in six similar motifs, each of which contains four β strands that form the blades of a propeller-like structure.

The active site is in the middle of one side of the propeller

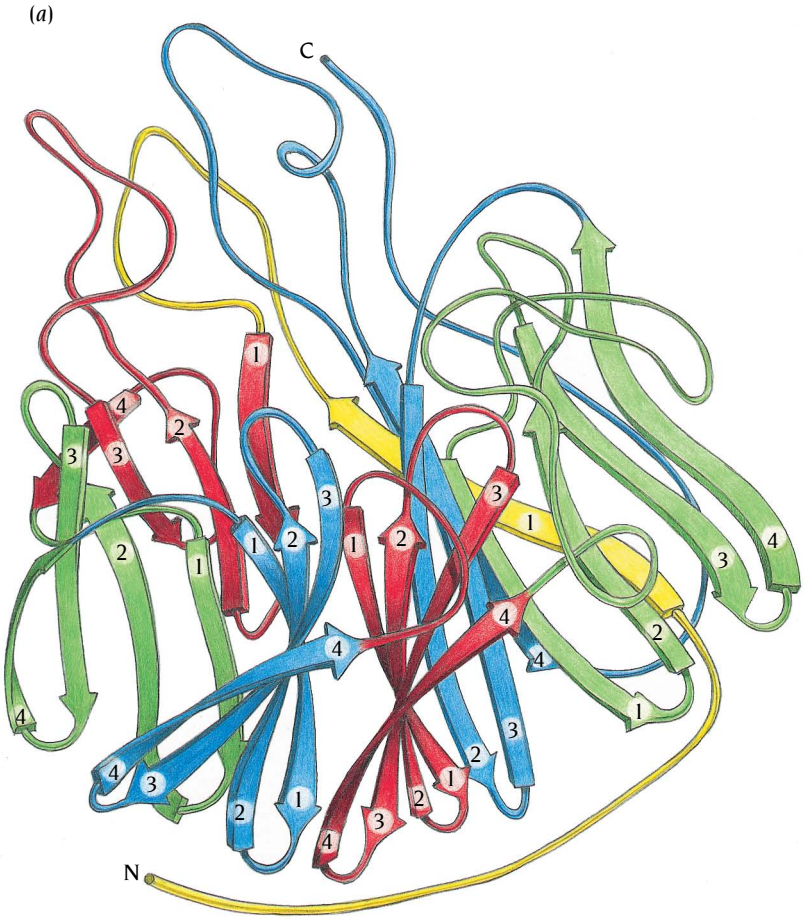
Not only are the topologies within the six β sheets in each subunit identical, but so are their connections to each other, with the exception of the last β sheet (see Figure 5.7b). The fourth strand of each β sheet is connected across the top of the subunit (seen in Figure 5.7a coming out of the page) to the first strand of the next sheet. The loop that connects strands 2 and 3 within the sheet is also at the top of the subunit.

Furthermore, because of the approximate sixfold symmetry of the β-sheet motifs, these 12-loop regions, derived from the six β sheets, are on the same side of the molecule, as can be seen in Figure 5.9a, where we see a single polypeptide chain (one of the four subunits) from the side of the propeller. The β sheets are arranged cyclically around an axis through the center of the molecule. The loop regions at the top of this barrel are extensive (Figure 5.9a) and together they form a wide funnel-shaped pocket containing the active site (Figure 5.9b). This is analogous to the active site formed by the loop regions at the top of the α/β-barrel structures.

Greek key motifs occur frequently in antiparallel β structures

We saw in Chapter 2 that the **Greek key motif** provides a simple way to connect antiparallel β strands that are on opposite sides of a barrel structure. We will now look at how this motif is incorporated into some of the simple antiparallel β-barrel structures and show that an antiparallel β sheet of eight strands can be built up only by hairpin and/or Greek key motifs, if the connections do not cross between the two ends of the β sheet.

Petitioner Merck, Ex. 1014, p. 72



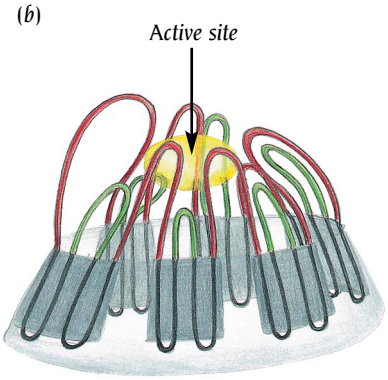


Figure 5.9 The six four-stranded motifs in a single subunit of neuraminidase form the six blades of a propeller-like structure. A schematic diagram of the subunit structure shows the propeller viewed from its side (a). An idealized propeller structure viewed from the side to highlight the position of the active site is shown in (b). The loop regions that connect the motifs (red in b) in combination with the loops that connect strands 2 and 3 within the motifs (green in b) form a wide funnel-shaped active site pocket. [(a) Adapted from P. Colman et al., *Nature* 326: 358–363, 1987.]

Assume that we have eight antiparallel β strands arranged in a barrel structure. We decide that we want to connect strand number n to an antiparallel strand at the same end of the barrel. We do not want to connect it to strand number $n + 1$ as in the up-and-down barrels just described, nor do we want to connect it to strand number $n - 1$ which is equivalent to turning the up-and-down barrel in Figure 5.2 upside down. What alternatives remain?

It is easy to see from Figure 5.10 that there are only two alternatives. We can connect it either to strand number $n + 3$ or to $n - 3$. Both cases require only short loop regions that traverse the end of the barrel. How do we now continue the connections? The simplest way to connect the strands that were skipped over is to join them by up-and-down connections, as illustrated in Figure 5.10.

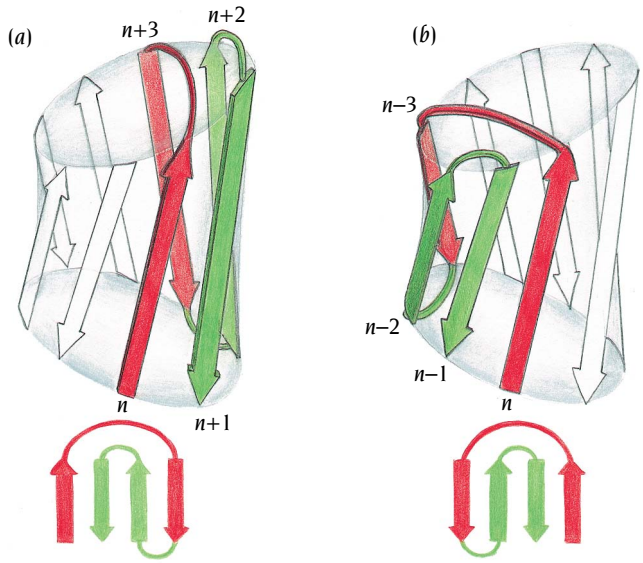


Figure 5.10 Idealized diagrams of the Greek key motif. This motif is formed when one of the connections of four antiparallel β strands is not a hairpin connection. The motif occurs when strand number n is connected to strand $n + 3$ (a) or $n - 3$ (b) instead of $n + 1$ or $n - 1$ in an eight-stranded antiparallel β sheet or barrel. The two different possible connections give two different hands of the Greek key motif. In all protein structures known so far only the hand shown in (a) has been observed.

Petitioner Merck, Ex. 1014, p. 73

We have now connected four adjacent strands of the barrel in a simple and logical fashion requiring only short loop regions. The result is the Greek key motif described in Chapter 2, which is found in the large majority of antiparallel β structures. The two cases represent the two possible different hands, but in all structures known to us the hand that corresponds to the case where β strand *n* is linked to β strand *n* + 3 as in Figure 5.10a is present.

The remaining four strands of the barrel can be joined either by up-and-down connections before and after the motif or by another Greek key motif. We will examine examples of both cases.

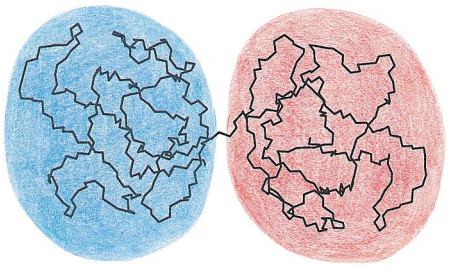


Figure 5.11 A computer-generated diagram of the structure of γcrystallin comprising one polypeptide chain of 170 amino acid residues. The diagram illustrates that the polypeptide chain is arranged in two domains (blue and red). Only main chain (N, C′, C_α) atoms and no side chains are shown.

The γ–crystallin molecule has two domains

The transparency and refractive power of the lenses of our eyes depend on a smooth gradient of refractive index for visible light. This is achieved partly by a regular packing arrangement of the cells in the lens and partly by a smoothly changing concentration gradient of lens-specific proteins, the crystallins.

There are at least three different classes of crystallins. The α and β are heterogeneous assemblies of different subunits specified by different genes, whereas the **gamma (γ) crystallins** are monomeric proteins with a polypeptide chain of around 170 amino acid residues. The structure of one such γ crystallin was determined in the laboratory of Tom Blundell in London to 1.9 Å resolution. A picture of this molecule generated from a graphics display is shown in Figure 5.11.

Let us now examine this molecule and dissect it into its structural components to see if we can understand how these are put together. We will reduce this rather complex, and at first sight bewildering, structure to its simplest representation as a series of motifs. This will help us to understand the structure and see its relationships to other structures.

We can immediately discern from Figure 5.11 that the molecule is divided into two clearly separated domains that seem to be of similar size. For the next step we would need a stereopicture of the model or, much better, a graphics display where we could manipulate the model and look at it from different viewpoints. Here instead we have made a schematic diagram of one domain (Figure 5.12), which is normally not done until the analysis is completed and the structural principles are clear.

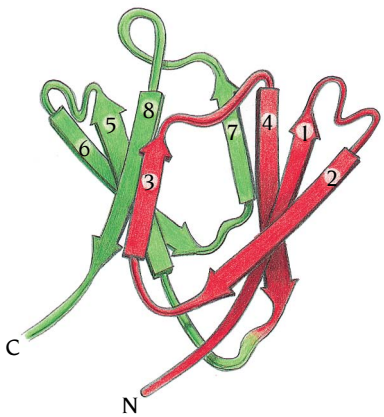


Figure 5.12 Schematic diagram of the path of the polypeptide chain in one domain (the blue region in Figure 5.11) of the γ-crystallin molecule. The domain structure is built up from two β sheets of four antiparallel β strands, sheet 1 from β strands 1, 2, 4, and 7 and sheet 2 from strands 3, 5, 6, and 8.

The domain structure has a simple topology

We will now follow the main polypeptide chain and trace out a topological diagram for this domain. We can immediately see from Figure 5.13 that the only secondary structure in the molecule is made up of β strands, which are arranged in an antiparallel fashion into two separate β sheets. Beta strands 1, 2, 4, and 7 form one antiparallel β sheet with the strand order 2 1 4 7. We thus draw the left four arrows in Figure 5.13 and connect strands 1 and 2. Similarly, we see that β strands 3, 5, 6, and 8 form another antiparallel β sheet with the strand order 6 5 8 3. We notice that strands 7 and 6 are adjacent although not hydrogen bonded to each other on the back side of the domain. We thus position strand 6 adjacent to strand 7 in the topology diagram but make a space between them to indicate that they belong to different β sheets. Alternatively, we could have positioned strands 2 and 3 adjacent to each other, which would have given a topologically identical diagram. We then connect the strands in consecutive order along the polypeptide chain.

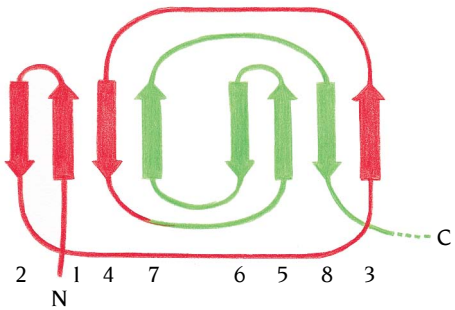


Figure 5.13 A preliminary topological diagram of the structure of one domain of γ crystallin shown in Figure 5.12, illustrating that the two β sheets are separate within the domain.

Two Greek key motifs form the domain

The topological diagram of Figure 5.13 has been drawn to reflect the observation that the two β sheets are separate: β strands 2 and 3 are not hydrogen bonded to each other, nor are strands 6 and 7. The connections

Petitioner Merck, Ex. 1014, p. 74

look unnecessarily complicated, but notice from the schematic diagram of the domain in Figure 5.12 that the two β sheets are packed against each other so that they form a distorted barrel. To see if the diagram can be simplified, we idealize the barrel and plot the strands along the surface of the barrel as shown in Figure 5.14. It is then immediately obvious that strands 1, 2, 3, and 4 form a Greek key motif, as do strands 5, 6, 7, and 8. These two motifs are joined by a loop across the bottom of the barrel, between strands 4 and 5.

On the basis of this new insight we can draw the topology diagram shown in the left half of Figure 5.15b. What is the difference between this and the previous topological diagram we made? The only changes we have made are to move β strand 3 from the right edge to the left edge of the domain topology and to close the gap between strands 7 and 6. We have changed neither the strand order nor the connections between the strands; thus the two diagrams are topologically identical.

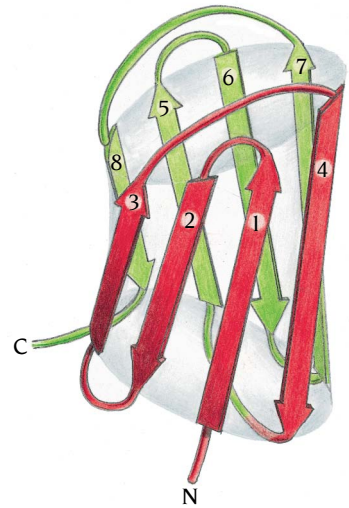


Figure 5.14 The eight β strands in one domain of the crystallin structure in this idealized diagram are drawn along the surface of a barrel. From this diagram it is obvious that the β strands are arranged in two Greek key motifs, one (red) formed by strands 1–4 and the other (green) by strands 5–8. Notice that the β strands that form one motif contribute to both β sheets as shown in Figure 5.12.

The two domains have identical topology

Using a graphics display, we could do the same thing for the second domain and arrive at the full topology diagram in Figure 5.15b. From this diagram it is obvious that the two domains have identical topology and thus in all probability similar structures. This realization is not at all trivial. To be able

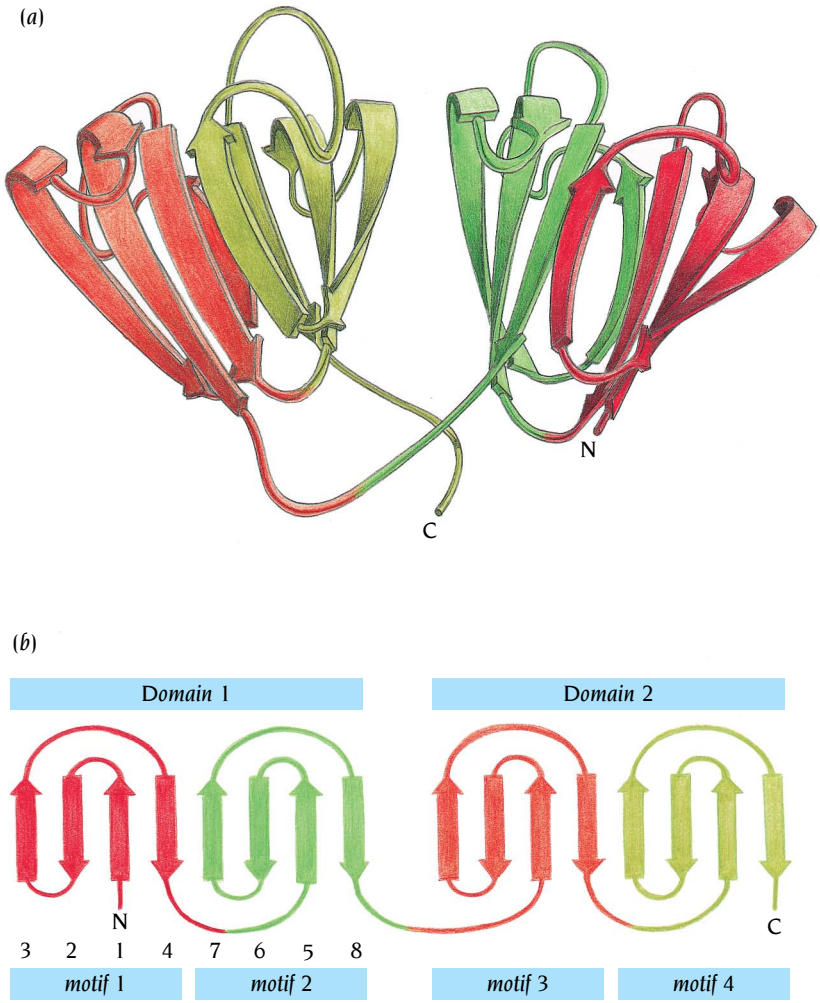


Figure 5.15 Schematic diagram (a) and topology diagram (b) for the γ-crystallin molecule. The two domains of the complete molecule have the same topology; each is composed of two Greek key motifs that are joined by a short loop region. [(a) Adapted from T. Blundell et al., *Nature* 289: 771–777, 1981.]

Petitioner Merck, Ex. 1014, p. 75

to see it when one looks at the structure on the display requires considerable experience because the two domains are in different orientations in the molecule. The brain therefore has to store the image of one domain while examining different orientations of the second domain. A topology diagram, on the other hand, immediately reveals similarities in domain structures. This illustrates one very important use of topology diagrams—namely, to reduce a complicated pattern to a simpler one, from which conclusions can be drawn that are also valid for the complicated pattern.

The two domains have similar structures

A relevant question to ask at this stage is, do the topological identities displayed in the diagram reflect structural similarity? We can now see that topologically the polypeptide chain is divided into four consecutive Greek key motifs arranged in two domains. How similar are the domain structures to each other, and how similar are the two motifs within each domain?

Tom Blundell has answered these questions by superposing the C_α atoms of the two motifs within a domain with each other and by superposing the C_α atoms of the two domains with each other. As a rule of thumb, when two structures superpose with a mean deviation of less than 2 Å they are considered structurally equivalent. For each pair of motifs Blundell found that 40 C_α atoms superpose with a mean distance of 1.4 Å. These 40 C_α atoms within each motif are therefore structurally equivalent. Since each motif comprises only 43 or 44 amino acid residues in total, these comparisons show that the structures of the complete motifs are very similar. Not only are the individual motifs similar in structure, but they are also pairwise arranged into the two domains in a similar way since superposition of the two domains showed that about 80 C_α atoms of each domain were structurally equivalent.

This structural similarity is also reflected in the amino acid sequences of the domains, which show 40% identity. They are thus clearly homologous to each other. The motif structures within the domains superpose equally well but their sequence homology is less, being around 30% between motifs 1 and 2 and 20% between 3 and 4. This study, however, clearly shows that the topological description in terms of four Greek key motifs is also valid at the structural and amino acid sequence levels.

The Greek key motifs in γ crystallin are evolutionarily related

These comparisons strongly suggest that the four Greek key motifs are evolutionarily related. We can guess from the amino acid sequence comparison that this protein evolved in two stages, beginning with the duplication of a primordial gene coding for one motif of about 40 amino acid residues, followed by fusion of the duplicated genes to give a single gene encoding one domain. The gene for this domain, we may imagine, later duplicated in turn and fused to give the full gene for the present-day γ-crystallin polypeptide. The evidence that this was the second step lies in the fact that the amino acid sequence homology is greater between the domains than between the motifs within each domain.

There is some circumstantial evidence in the organization of the crystallin gene for the evolutionary history that we have reconstructed. The amino acid sequence of a mouse β crystallin is homologous to that of γ crystallin and shows the same four homologous motifs. Its coding sequence is in separate DNA sequences (exons) interrupted by noncoding DNA sequences (introns). Walter Gilbert at Harvard University suggested in 1978 that genes for large proteins might have evolved by the accidental juxtaposition of exons coding for specific functions. In β crystallin the three introns are positioned at the junctions between the four motifs, supporting Gilbert's ideas. These introns could, therefore, be evolutionary remnants of the gene duplication and fusion events.

Petitioner Merck, Ex. 1014, p. 76

The Greek key motifs can form jelly roll barrels

In antiparallel barrel structures with the Greek key motif one of the connections in the motif is made across one end of the barrel. Such connections can be made several times in a β barrel giving different variations and combinations of the Greek key motif. In the structure of crystallin there are two consecutive Greek key motifs that form a barrel with two such connections. There is a different but frequently occurring motif, the **jelly roll motif**, in which there are four connections of this type. It is called jelly roll, or Swiss roll, because the polypeptide chain is wrapped around a barrel core like a jelly roll. This motif has been found in a variety of different structures including the coat proteins of most of the spherical viruses examined thus far by x-ray crystallography, the plant lectin concanavalin A and the hemagglutinin protein from the influenza virus.

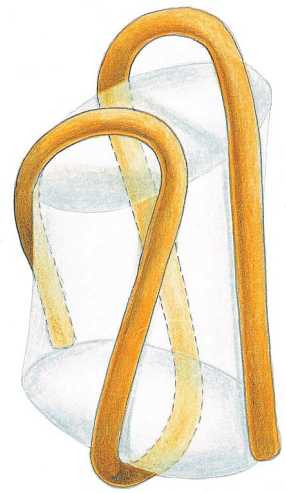


Figure 5.16 A diagram of a piece of string wrapped around a barrel to illustrate the basic pattern of a jelly roll motif.

The jelly roll motif is wrapped around a barrel

To illustrate how this rather complicated structure is built up, we will start by wrapping a piece of string around a barrel as shown in Figure 5.16. The string goes up and down the barrel four times, crosses over once at the bottom and twice at the top of the barrel. This configuration is the basic pattern for the jelly roll motif.

Let us do the same thing with a strip of paper the width of which is approximately one-eighth of the circumference of the top of the barrel. We imagine that a polypeptide chain follows the edges of this strip, starting at the bottom right corner of the strip and ending at the bottom left corner (Figure 5.17a). The polypeptide chain has eight straight sections, β strands, interrupted by loop regions. The β strands are arranged in a long antiparallel hairpin such that strand 1 is hydrogen bonded to strand 8, strand 2 to strand 7, and so on.

We now wrap the strip around the barrel following the path of the string in Figure 5.16 and in such a way that the β strands go along the sides of the barrel and the loop regions form the connections at the top and bottom of the barrel (Figure 5.17b).

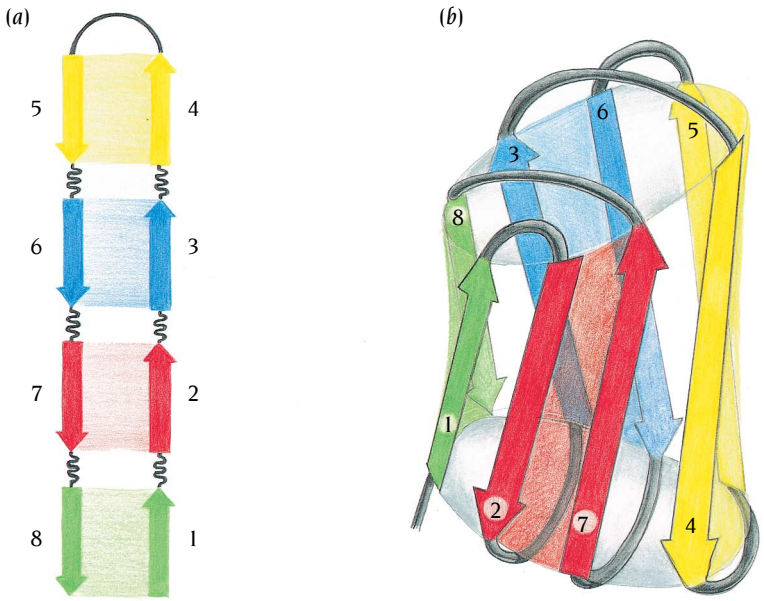


Figure 5.17 A simple illustration of the way eight β strands are arranged in a jelly roll motif. (a) The eight β strands are drawn as arrows along two edges of a strip of paper. The strands are arranged such that strand 1 is opposite strand 8, etc. The β strands are separated by loop regions. (b) The strip of paper in (a) is wrapped around a barrel in the same way as the string in Figure 5.16, such that the β strands follow the surface of the barrel and the loop regions (gray) provide the connections at both ends of the barrel. The β strands are now arranged in a jelly roll motif.

Petitioner Merck, Ex. 1014, p. 77

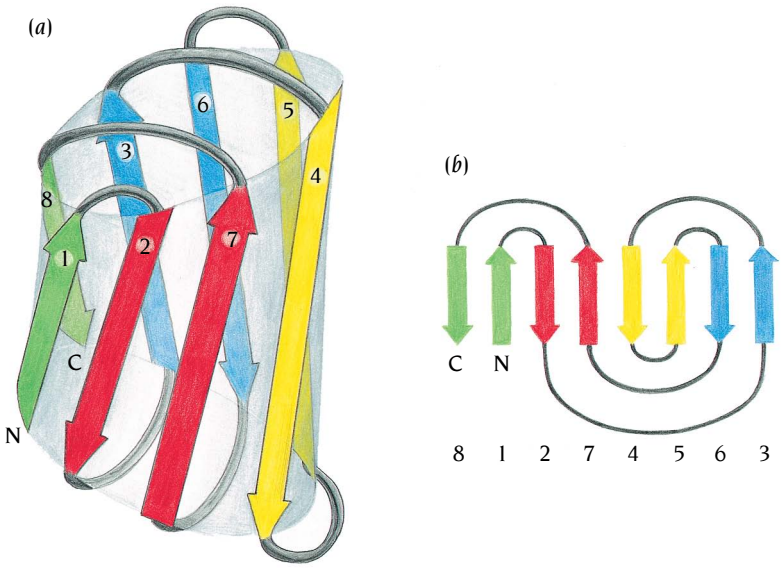


Figure 5.18 Topological diagrams of the jelly roll structure. The same color scheme is used as in Figure 5.17.

The hydrogen-bonded antiparallel β strand pairs 1:8, 2:7, 3:6, and 4:5 are now arranged such that β strand 1 is adjacent to strand 2, 7 is adjacent to 4, 5 to 6, and 3 to 8. These can also form hydrogen bonds to each other. All adjacent β strands are antiparallel. This is the basic jelly roll β-barrel structure for eight β strands (Figure 5.18a). Most such barrels have eight strands, but any even number of strands greater than four can form a jelly roll barrel. In eight-stranded barrels there are two connections across the top of the barrel and two across the bottom. In addition, there are two connections between adjacent β strands at the top and one at the bottom. A topological diagram of this fold is given in Figure 5.18b.

The jelly roll barrel is thus conceptually simple, but it can be quite puzzling if it is not considered in this way. Discussion of these structures will be exemplified in this chapter by hemagglutinin and in Chapter 16 by viral coat proteins.

The jelly roll barrel is usually divided into two sheets

The barrels we have used to illustrate both the Greek key and the jelly roll structures provide topological descriptions, as defined in Chapter 2. A topological description accurately represents the connectivity and the strand order around the barrel and thus is very useful in the same way that a subway map tells you how stations are interconnected. However, when one analyzes the pattern of hydrogen bonds between the β strands of such barrels, one finds that they usually form two sheets with few if any hydrogen bonds between strands that belong to the different β sheets, as we saw in the crystallin structure. The barrel is distorted and adjacent β strands are separated from each other in two places across the barrel. The division of β strands into these two sheets does not necessarily follow the division into topological motifs. The β strands in jelly roll barrels are also usually arranged in two sheets that are packed against each other. This does not, however, change either the topology or the usefulness of the description of these structures as barrels as long as one keeps in mind that these barrels are distorted and flattened.

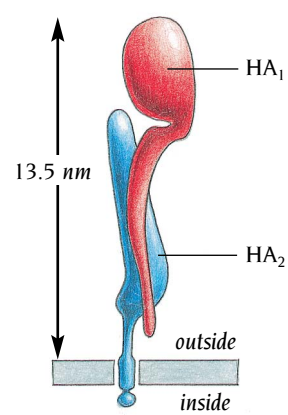


Figure 5.19 Schematic picture of a single subunit of influenza virus hemagglutinin. The two polypeptide chains HA_1 and HA_2 are held together by disulfide bridges.

Petitioner Merck, Ex. 1014, p. 78

The functional hemagglutinin subunit has two polypeptide chains

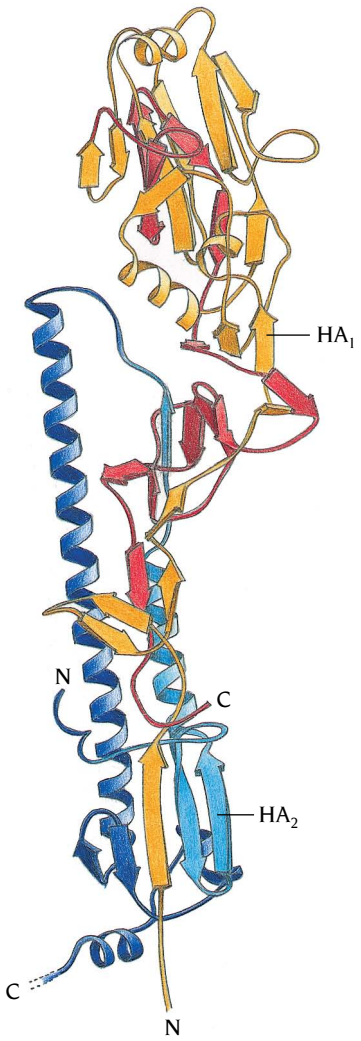
We have already discussed one envelope protein of influenza virus, neuraminidase, as an example of an up-and-down antiparallel β motif. In the second envelope protein, hemagglutinin, one domain of the polypeptide chain is folded into a jelly roll motif. We shall now look at some other features of hemagglutinin that are important for its biological function.

The hemagglutinin polypeptide chain is synthesized on membrane-bound ribosomes of the rough endoplasmic reticulum and then cotranslationally inserted into the membrane. The polypeptide chain is proteolytically cleaved to yield two chains of 328 and 221 amino acids called HA_1 and HA_2, respectively, which are held together by disulfide bonds (Figure 5.19). Three hemagglutinin monomers, each with one HA_1 and one HA_2 chain, trimerize in the rough endoplasmic reticulum and are transported from there through the Golgi apparatus to the plasma membrane, where the functional part of the molecule is now outside the cell anchored to the cell membrane via the HA_2 tails.

Progeny virus particles then bud from patches of the infected cell's plasma membrane that contain both the viral hemagglutinin and neuraminidase. The viral envelopes therefore contain both viral membrane proteins but no cellular membrane proteins.

Protease treatment of influenza virus particles cleaves the three HA_2 chains of the trimeric hemagglutinin molecules. Such cleavage leaves three C-terminal chains, 47 amino acids long, still inserted into the viral envelope and releases a second much larger soluble fragment. The soluble fragment consists of three complete HA_1 chains disulfide bonded to three HA_2 chains complete except for their membrane anchor regions. This soluble trimeric fragment has been crystallized at neutral pH and its structure was determined at 3 Å resolution and subsequently refined to a high resolution in the laboratory of Don Wiley at Harvard University. The influenza virus that he used was the Hong Kong 1968 strain, which caused the "Asian flu" pandemic disease.

Figure 5.20 Schematic diagram of the subunit structure of hemagglutinin from influenza virus. The structure comprises about 550 amino acids arranged in two chains HA_1 (red) and HA_2 (blue). The first half of each chain has a lighter color in the diagram. The subunit is very elongated with a long stemlike region built up by residues from both chains and includes one of the longest α helices known in a globular structure, about 75 Å long. The globular head is formed by residues only from HA_1. (Courtesy of Don Wiley, Harvard University.)



The subunit structure is divided into a stem and a tip

The monomeric subunit is divided into a long, fibrous stemlike region extending outward from the membrane with a globular region at its tip (Figure 5.20). The globular region contains only residues of HA_1, while the stem contains some residues of HA_1 and all of HA_2.

The amino terminus of HA_1 is found at the base of the stem close to the viral membrane. The first 63 amino acids of HA_1 (pale red in Figure 5.20) reach, in an almost fully extended structure, nearly 100 Å along the length of the molecule before the first compact fold. These 63 residues form part of the stem region of the subunit. The globular tip is an eight-stranded distorted jelly roll structure comprising residues 116 to 261, which are folded into a distorted barrel. The remaining 70 residues of HA_1 return to the stem region, running nearly antiparallel to the initial stretch of 63 residues.

The major structural feature of the HA_2 chain (blue in Figure 5.20) is a hairpin loop of two α helices packed together. The second α helix is 50 amino acids long and reaches back 76 Å toward the membrane. At the bottom of the stem there is a β sheet of five antiparallel strands. The central β strand is from HA_1, and this is flanked on both sides by hairpin loops from HA_2. About 20 residues at the amino terminal end of HA_2 are associated with the activity by which the virus penetrates the host cell membrane to initiate infection. This region, which is quite hydrophobic, is called the fusion peptide.

The hemagglutinin trimer molecule is 135 Å long (from membrane to tip) and varies in cross-section between 15 Å and 40 Å. It is thus an unusually

Petitioner Merck, Ex. 1014, p. 79

elongated molecule. The long fibrous stems of each subunit form the major subunit contacts (Figure 5.21). In particular, the three long HA_2 helices, one from each subunit, intertwine by forming a coiled-coil structure (see Chapter 3) and pack against each other for part of their lengths forming a core 40 Å long stabilized by both hydrophobic residues and internal salt bridges. In addition the three heads of the subunits are close together and interact with each other, further stabilizing the trimer.

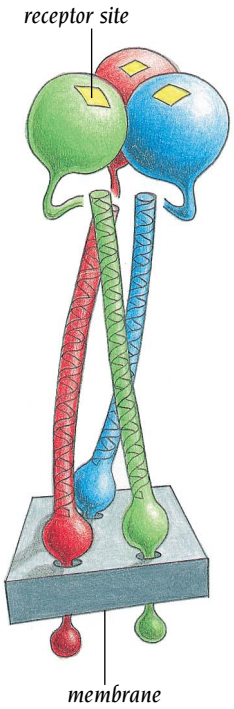


Figure 5.21 The hemagglutinin molecule is formed from three subunits. Each of these subunits is anchored in the membrane of the influenza virus. The globular heads contain the receptor sites that bind to sialic acid residues on the surface of eukaryotic cells. A major part of the subunit interface is formed by the three long intertwining helices, one from each subunit. (Adapted from I. Wilson et al., *Nature* 289: 366–373, 1981.)

The receptor binding site is formed by the jelly roll domain

To initiate infection, the virus hemagglutinin binds to the sialic acid residues of glycosylated receptor proteins on the target cell surface. Once bound to the receptor, the virus is then taken into the cell by endocytosis. The receptor binding site on the hemagglutinin molecule has been determined from experiments with mutants of hemagglutinin and from the determination of structures with bound inhibitors (modified sialic acid molecules with substituents at two different positions of the sugar ring, Figure 5.22).

The binding site is located at the tip of the subunit within the jelly roll structure (Figure 5.23). The sialic acid moiety of the hemagglutinin inhibitors binds in the center of a broad pocket on the surface of the barrel (Figure 5.24). In addition to this groove there is a hydrophobic channel that can accomodate large hydrophobic substituents at the C_2 position of sialic acid (Figures 5.22 and 5.24).

Antibodies in our immune system bind to the receptor binding site, so preventing the virus from entering a cell. The virus can escape this neutralization through mutations in residues that form this binding site. Such mutations, however, are found only at the rim of the sialic acid binding pocket, presumably because mutation of residues inside this pocket would prevent the virus from binding to the cell surface receptor and consequently prevent viral propagation. The receptor binding site is therefore an ideal target for drug design, and studies of inhibitor binding to this site have provided valuable clues and ideas for the design of molecules that are candidate drug therapies for influenza virus infections.

Hemagglutinin acts as a membrane fusogen

In addition to binding to sialic acid residues of the carbohydrate side chains of cellular proteins that the virus exploits as receptors, hemagglutinin has a second function in the infection of host cells. Viruses, bound to the plasma membrane via their membrane receptors, are taken into the cells by endocytosis. Proton pumps in the membrane of endocytic vesicles that now contain the bound viruses cause an accumulation of protons and a consequent lowering of the pH inside the vesicles. The acidic pH (below pH 6) allows hemagglutinin to fulfill its second role, namely, to act as a membrane fusogen by inducing the fusion of the viral envelope membrane with the membrane of the endosome. This expels the viral RNA into the cytoplasm, where it can begin to replicate.

This fusogenic activity of influenza hemagglutinin is frequently exploited in the laboratory. If, for example, the virus is bound to cells at a temperature too low for endocytosis and then the pH of the external medium is lowered, the hemagglutinin causes direct fusion of the viral envelope with the plasma membrane; infection is achieved without endocytosis. Similarly, artificial vesicles with hemagglutinin in their membrane and other molecules in their lumen can be caused to fuse with cells by first allowing the vesicles to bind to the plasma membrane via the hemagglutinin and then lowering the pH of the medium. In this way the contents of the vesicles are delivered to the recipient cell's cytoplasm.

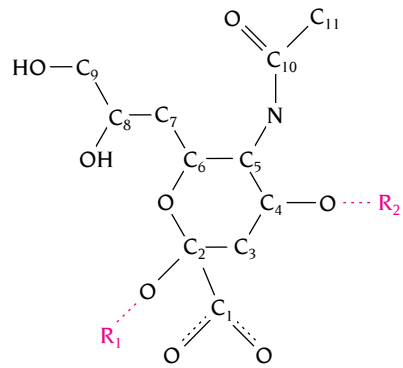
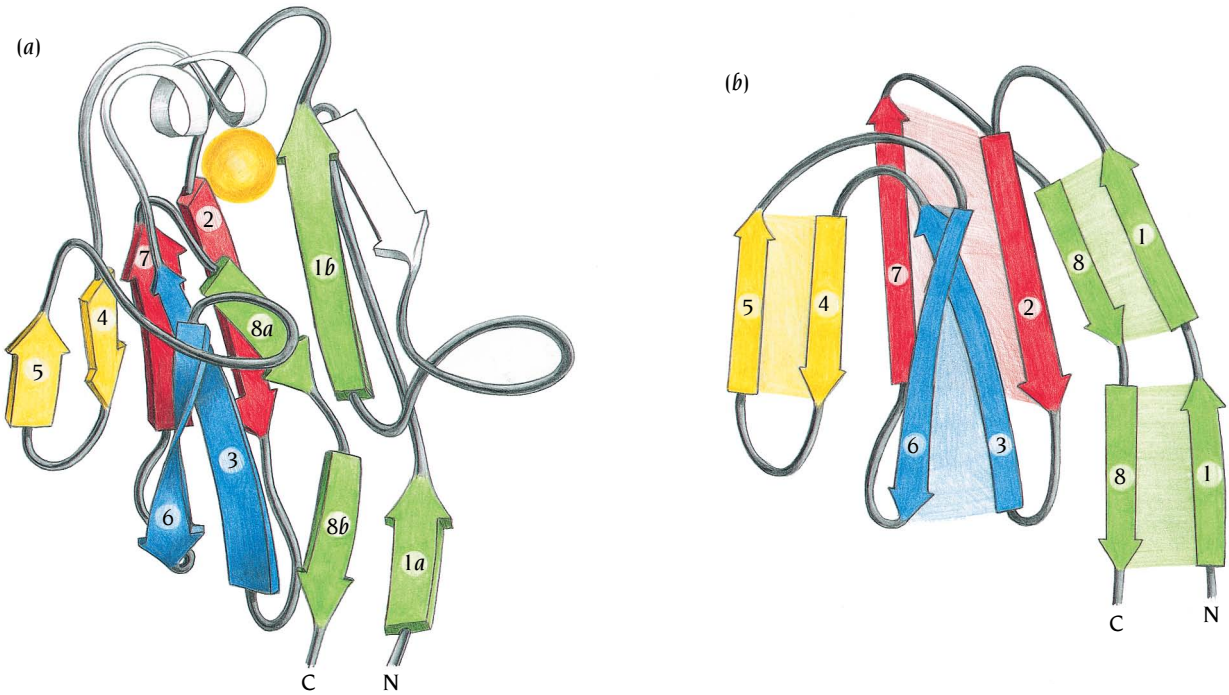


Figure 5.22 Chemical formula for sialic acid (α–5–*n*–acetylneuraminic acid) drawn in approximately the same orientation as the ball and stick models in Figure 5.24. R_1 and $R_{2'}$ which are H atoms in sialic acid, denote substituents introduced to design tightly bound inhibitors. These are large and hydrophobic as shown in Figure 5.24.

Petitioner Merck, Ex. 1014, p. 80



The structure of hemagglutinin is affected by pH changes

The structure of the soluble trimeric hemagglutinin fragment above pH 6 gave no indication of how hemagglutinin induces membrane fusion. When the soluble trimers were exposed to pH below 6 they aggregated and therefore could not be crystallized. However, by using monoclonal antibodies to specific epitopes on the two chains, HA_1 and HA_2, it was shown that lowering the pH causes a massive conformational change in hemagglutinin. As a result of this induced conformational change the hemagglutinin becomes highly susceptible to proteolytic cleavage. This property was used in the laboratory of Don Wiley to produce a soluble fragment of the low pH form comprising

Figure 5.23 The globular head of the hemagglutinin subunit is a distorted jelly roll structure (a). β strand 1 contains a long insertion, and β strand 8 contains a bulge in the corresponding position. Each of these two strands is therefore subdivided into shorter β strands. The loop region between β strands 3 and 4 contains a short α helix, which forms one side of the receptor binding site (yellow circle). A schematic diagram (b) illustrates the organization of the β strands into a jelly roll motif.

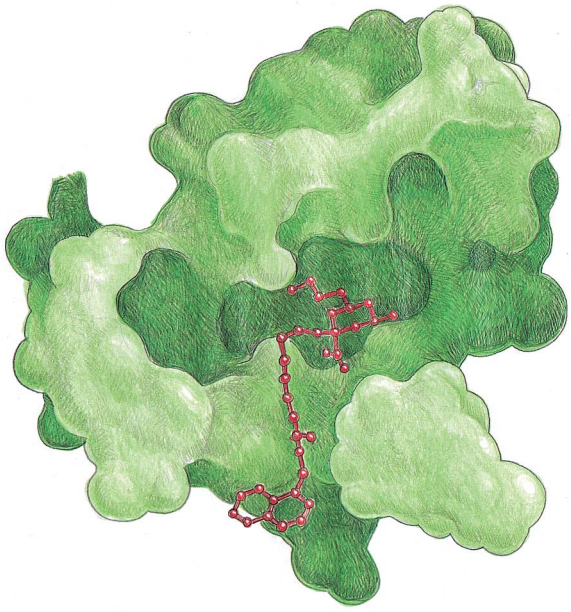


Figure 5.24 Space-filling model (green) of the sialic acid binding domain of hemagglutinin with a bound inhibitor (red) illustrating the different binding grooves. The sialic acid moiety of the inhibitor binds in the central groove. A large hydrophobic substituent, R_1, at the C_2 position of sialic acid binds in a hydrophobic channel that runs from the central groove to the bottom of the domain. (Adapted from S.J. Watowich et al., *Structure* 2: 719–731, 1994.)

Petitioner Merck, Ex. 1014, p. 81

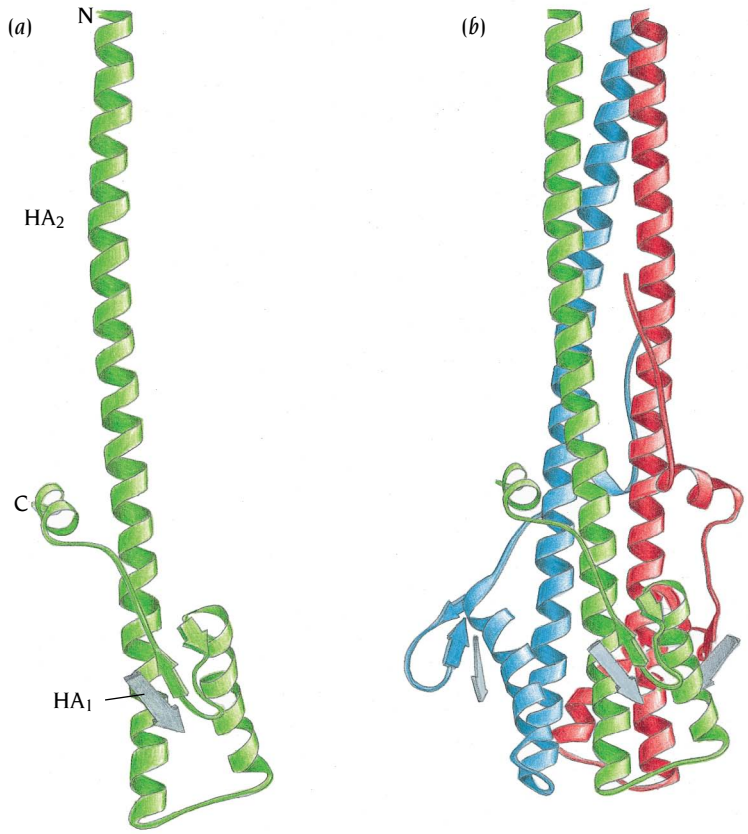


Figure 5.25 Schematic diagrams of the structure of a proteolytic fragment of hemagglutinin at low pH where the molecule induces membrane fusion. (a) Residues 38–175 of the HA_2 polypeptide chain (green) form a 100-Å long α helix starting at the N-terminus followed by a loop, a β hairpin and finally a short C-terminal helix. In addition the diagram shows a β strand from the HA_1 polypeptide chain (gray) which participates with the β hairpin to form a three-stranded antiparallel β sheet. (b) The proteolytic fragment forms a trimer like the intact hemagglutinin molecule. The three long α helices of the three subunits intertwine to form a three-stranded coiled coil. (Adapted from P.A. Bullough et al., *Nature* 371: 37–43, 1994.)

residues 1–27 of the HA_1 chain and residues 38–175 of the HA_2 chain that could be crystallized; its structure was determined in 1994 (Figure 5.25).

The structure of this fragment, which like the high pH fragment is a trimer, confirms that the HA_2 subunit undergoes major structural changes at low pH. Most of the secondary structure elements are essentially preserved but there are two important exceptions (Figure 5.26). First, the loop region B between the two long α helices A and C + D in the high pH structure changes into an α helix. Second, an α-helical region in the middle of helix C + D changes into a loop region (Figure 5.26). The resulting helix A + B + C comprises 65 residues and is about 100 Å long. Previous examination of the amino acid sequence of this whole region, residues 40–106, had shown a very clear heptad repeat typical of coiled-coil structures (see Chapter 3). It was therefore surprising that in the high pH structure only a part of this region, helix C, actually has a coiled-coil structure. Reassuringly, in the low pH structure the whole region is a coiled coil which is involved in trimerization (Figure 5.25).

The new loop region between helices C and D changes completely the position and orientation of helix D so that instead of being continuous with helix C it is packed sideways against helix C in an antiparallel fashion (Figure 5.26). The small α helices and β strands at the C-terminal region of the chain collectively follow the realignment of helix D and occupy totally different positions (Figure 5.26).

The consequences for a possible fusion mechanism of this structural change are significant. In the high pH structure, the fusion peptide is attached to the N-terminus of helix A and is about 100 Å away from the receptor binding site. At low pH the N-terminus of helix A moves about 100 Å and is brought to the same region of the molecule as the receptor binding site (Figure 5.27). Even though the picture is incomplete, since only a proteolytic fragment of the low pH form was studied, it is quite clear that after these structural changes the likely positions of the receptor binding site and the fusion peptide are compatible with a fusion mechanism whereby hemagglutinin brings the viral and cellular membranes close together.

Petitioner Merck, Ex. 1014, p. 82

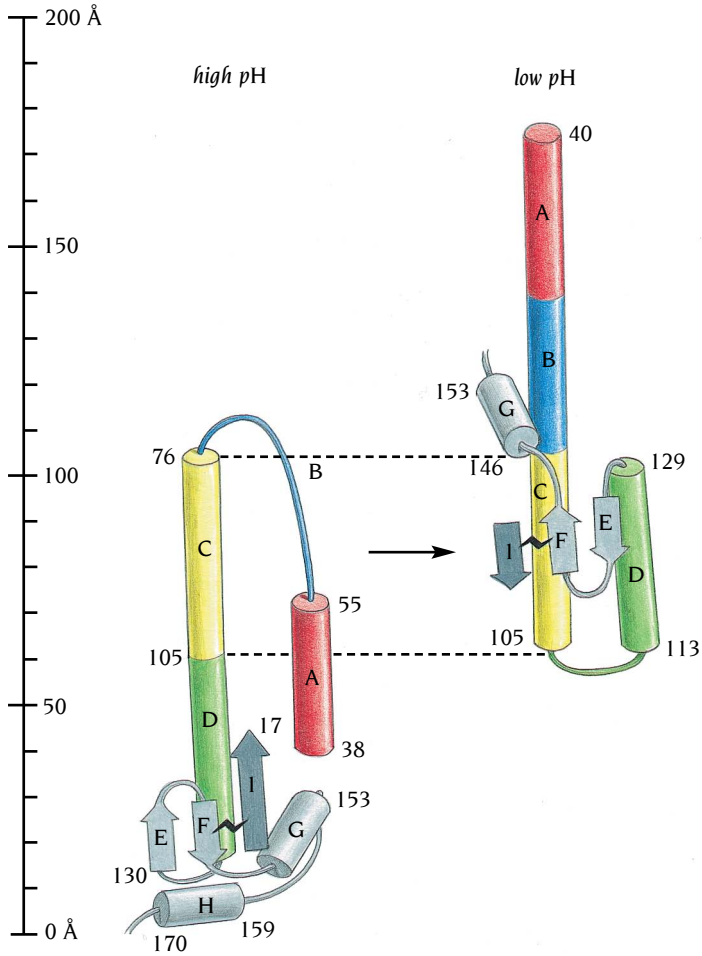


Figure 5.26 Schematic diagram illustrating the large conformational differences between the high and low pH forms of hemagglutinin. The loop region B (blue) in the high pH form has changed into an α helix producing a continuous 100-Å-long helix composed of regions A (red), B (blue) and C (yellow) at low pH. Furthermore, residues 105–113, which in the high pH form are in the middle of helix C–D, form a loop in the low pH form causing helix D (green) to be at a very different position. Consequently the β-hairpin E–F and the C-terminal helix G as well as the β strand I from HA_1 occupy very different positions in the two forms even though they have the same internal structure. The squiggle between β strands F and I denotes the S-S bond that joins HA_1 to HA_2. (Adapted from P.A. Bullough et al., *Nature* 371: 37–43, 1994.)

The large conformational change induced by low pH is irreversible. The low pH form is more thermostable than the high pH form and does not revert to the initial state when pH is raised. This suggests that the low pH form is lower in energy than the high pH form and that the energy required for fusion is stored up during the formation of the hemagglutinin molecule. Each of the three subunits of the trimeric hemagglutinin starts out as a single

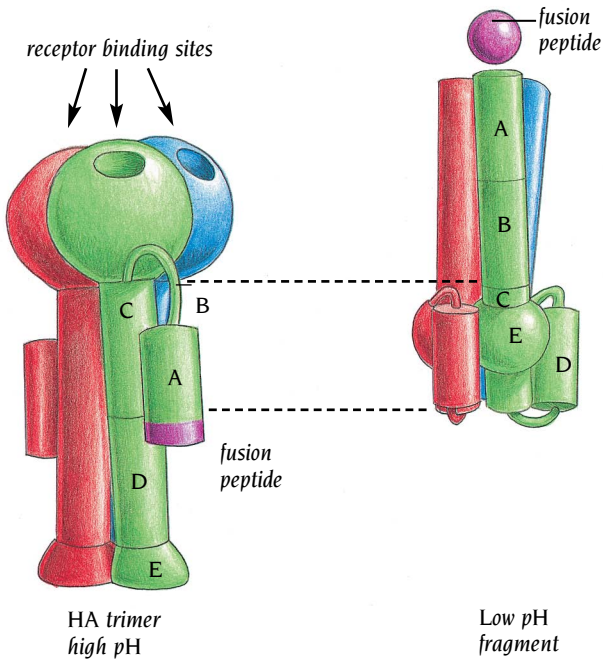


Figure 5.27 Schematic representation of a model for the conformational change of hemagglutinin that at low pH brings the fusion peptide to the same end of the molecule as the receptor binding site. The fusion peptide (purple) is at the end of helix A about 100 Å away from the receptor binding site in the high pH form. In the low pH fragment this region of helix A has moved about 100 Å towards the area where the receptor binding sites are expected to be in the intact hemagglutinin molecule. (Adapted from D. Stuart, *Nature* 371: 19–20, 1994.)

Petitioner Merck, Ex. 1014, p. 83

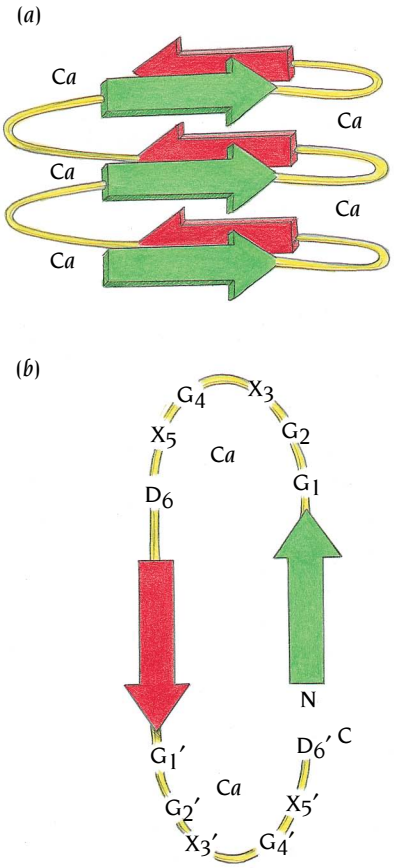


Figure 5.28 Schematic diagrams of the two-sheet β helix. Three complete coils of the helix are shown in (a). The two parallel β sheets are colored green and red, the loop regions that connect the β strands are yellow. (b) Each structural unit is composed of 18 residues forming a β-loop-β-loop structure. Each loop region contains six residues of sequence Gly-Gly-X-Gly-X-Asp where X is any residue. Calcium ions are bound to both loop regions. (Adapted from F. Jurnak et al., *Curr. Opin. Struct. Biol.* 4: 802–806, 1994.)

polypeptide chain, folded into a stable conformation. On cleavage to form the HA_1 and HA_2 chains of the mature molecules the free ends snap 20 Å apart to give the metastable high pH structure. This molecule is like a set trap: lowering the endosomal pH springs the trap, setting in motion a series of events which starts with a large conformational change to bring the fusion peptide into a position where it can engage the target membrane.

Parallel β–helix domains have a novel fold

In the first edition of this book this chapter was entitled "Antiparallel Beta Structures" but we have had to change this because an entirely unexpected structure, the β helix, was discovered in 1993. The β helix, which is not related to the numerous antiparallel β structures discussed so far, was first seen in the bacterial enzyme pectate lyase, the structure of which was determined by the group of Frances Jurnak at the University of California, Riverside. Subsequently several other protein structures have been found to contain β helices, including extracellular bacterial proteinases and the bacteriophage P22 tailspike protein.

In these β-helix structures the polypeptide chain is coiled into a wide helix, formed by β strands separated by loop regions. In the simplest form, the two-sheet β helix, each turn of the helix comprises two β strands and two loop regions (Figure 5.28). This structural unit is repeated three times in extracellular bacterial proteinases to form a right-handed coiled structure which comprises two adjacent three-stranded parallel β sheets with a hydrophobic core in between.

This structural organization has striking similarities to that of α/β proteins, the difference being that the loop-α helix-loop that connects the parallel β strands in α/β structures is substituted by a loop-β strand-loop in these β-helix structures. Instead of having one β sheet adjacent to a stack of α helices as in the α/β structures described in the previous chapter, β-helix structures have two parallel β sheets. In α/β structures a twist of about 20° between adjacent β strands is imposed by the packing requirements of the α helices: in order to pack ridges into grooves, as described in Chapter 3, the α helices have to be twisted with respect to each other and this forces the β strands also to be twisted. In β-helix structures no such constraint is present and therefore the sheets are almost planar and form straight walls (Figure 5.28a).

The basic structural unit of these two-sheet β helix structures contains 18 amino acids, three in each β strand and six in each loop. A specific amino acid sequence pattern identifies this unit; namely a double repeat of a nine-residue consensus sequence Gly-Gly-X-Gly-X-Asp-X-U-X where X is any amino acid and U is large, hydrophobic and frequently leucine. The first six residues form the loop and the last three form a β strand with the side chain of U involved in the hydrophobic packing of the two β sheets. The loops are stabilized by calcium ions which bind to the Asp residue (Figure 5.28). This sequence pattern can be used to search for possible two-sheet β structures in databases of amino acid sequences of proteins of unknown structure.

A more complex β helix is present in pectate lyase and the bacteriophage P22 tailspike protein. In these β helices each turn of the helix contains three short β strands, each with three to five residues, connected by loop regions. The β helix therefore comprises three parallel β sheets roughly arranged as the three sides of a prism. However, the cross-section of the β helix is not quite triangular because of the arrangement of the β sheets. Two of the sheets are

Petitioner Merck, Ex. 1014, p. 84

arranged adjacent to each other as in the two-sheet β helix, and the third β sheet is almost perpendicular to the other two (Figure 5.29). The β strands in these three parallel sheets are connected by three loop regions. One loop (loop *a* in Figure 5.29) is short and always formed by only two residues which have invariant conformations. The other two loops are much longer and vary in size and conformation. These long loops protrude from the β sheets and probably form the active site regions on the external surface of the protein. Since the long loop regions vary in size, the number and type of amino acids in each turn of the β-helix structures vary. Consequently, no specific amino acid sequence pattern has been identified for the three-sheet β-helix structures.

The number of helical turns in these structures is larger than those found so far in two-sheet β helices. The pectate lyase β helix consists of seven complete turns and is 34 Å long and 17–27 Å in diameter (Figure 5.30) while the β-helix part of the bacteriophage P22 tailspike protein has 13 complete turns. Both these proteins have other structural elements in addition to the β-helix moiety. The complete tailspike protein contains three intertwined, identical subunits each with the three-sheet β helix and is about 200 Å long and 60 Å wide. Six of these trimers are attached to each phage at the base of the icosahedral capsid.

The interior of the β-helix structure in pectate lyase is completely filled with side chains leaving no room for a channel. Interestingly, these side chains are not limited to hydrophobic groups but include polar and charged groups which are all neutralized either by hydrogen bonding or by electrostatic interactions. All the side chains in the interior of the helix are stacked such that the side chains of adjacent turns of the helix have a linear arrangement parallel to the helix axis. These internal stacks fall into different classes; polar stacks of Asn or Ser side chains, aliphatic stacks of Ala, Val, Leu or Ile side chains and aromatic stacks of Phe or Tyr side chains. Pectate lyase is an unusually stable protein and this stacking arrangement no doubt contributes to its stability.

Conclusion

Antiparallel β structures comprise the second major class of protein conformations. In these the antiparallel β strands are usually arranged in two β sheets that pack against each other and form a distorted barrel structure, the core of the molecule. Depending on the way the β strands around the barrel are connected along the polypeptide chain—in other words, depending on the topology of the barrels—they are divided into three main groups: up-and-down barrels, Greek key barrels, and jelly roll barrels.

The number of possible ways to form antiparallel β structures is very large. The number of topologies actually observed is small, and most β structures fall into these three major groups of barrel structures. The last two groups—the Greek key and jelly roll barrels—include proteins of quite diverse function, where functional variability is achieved by differences in the loop regions that connect the β strands that build up the common core region.

Up-and-down barrels are the simplest structures. Each β strand is connected to the next strand by a short loop region. Eight β strands arranged

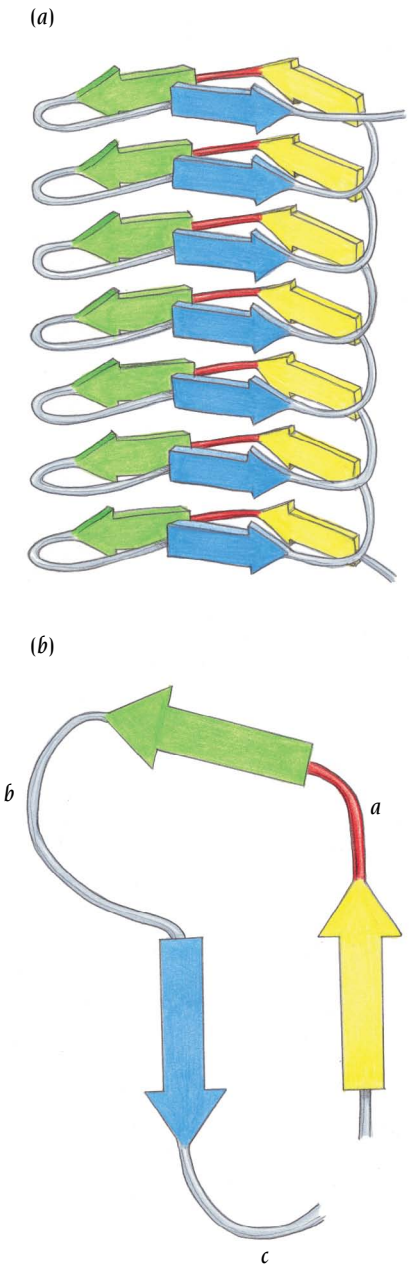


Figure 5.29 Schematic diagrams of the three-sheet β helix. (a) The three sheets of parallel β strands are colored green, blue and yellow. Seven complete coils are shown in this diagram but the number of coils varies in different structures. Two of the β sheets (blue and yellow) are parallel to each other and are perpendicular to the third (green). (b) Each structural unit is composed of three β strands connected by three loop regions (labeled *a*, *b* and *c*). Loop *a* (red) is invariably composed of only two residues, whereas the other two loop regions vary in length. (Adapted from F. Jurnak et al., *Curr. Opin. Struct. Biol.* 4: 802–806, 1994.)

Petitioner Merck, Ex. 1014, p. 85

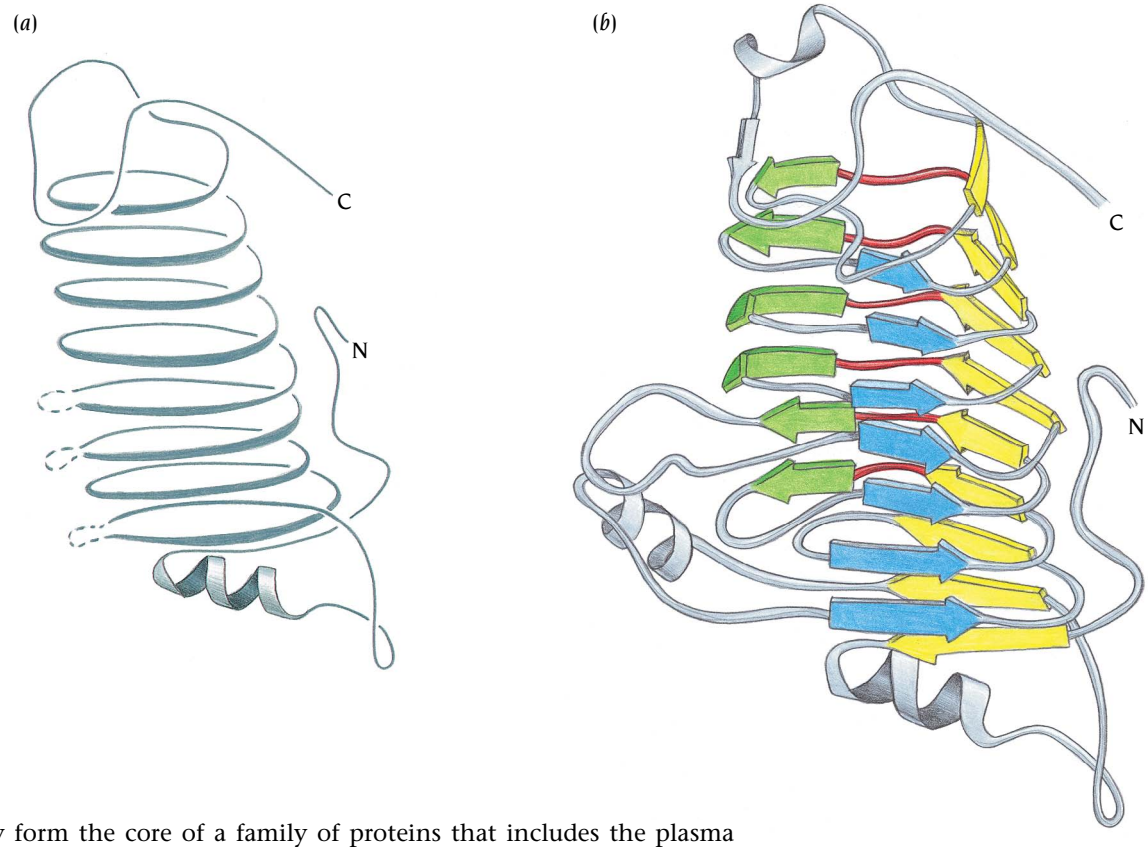


Figure 5.30 Schematic diagrams of the structure of the enzyme pectate lyase C, which has a three-sheet parallel β-helix topology. (a) Idealized diagram highlighting the helical nature of the path of the polypeptide chain which comprises eight helical turns. Dotted regions indicate positions where large external loops have been removed for clarity. (b) Ribbon diagram of the polypeptide chain. The predominant secondary structural elements are three parallel β sheets which are colored green, blue and yellow. Each β sheet is composed of 7–10 parallel β strands with an average length of four to five residues in each strand. The short loop regions of two residues length are shown in red. (Adapted from M.D. Yoder et al., *Science* 260: 1503–1507, 1993.)

in this way form the core of a family of proteins that includes the plasma retinol-binding protein in mammals, biliverdin-binding proteins in insects, and β-lactoglobulin from milk. Members of this family, as well as of the related P2 family with 10 β strands, bind large hydrophobic ligands inside the barrel. The barrel seems to be particularly suited to act as a container for chemically quite diverse ligands. Diversity in ligand binding is achieved by differences in the size of the barrel and in the amino acids that also participate in building up the common core.

Most of the known antiparallel β structures, including the immunoglobulins and a number of different enzymes, have barrels that comprise at least one Greek key motif. An example is γ crystallin, which has two consecutive Greek key motifs in each of two barrel domains. These four motifs are homologous in terms of both their three-dimensional structure and amino acid sequence and are thus evolutionarily related.

The jelly roll barrels are found in a variety of protein molecules, including viral coat proteins and hemagglutinin from influenza virus. This structure looks complicated but, in principle, is very simple if one thinks of the analogy of wrapping a strip of paper around a barrel, like a jelly roll. The hemagglutinin receptor-binding domain forms such a jelly roll barrel of eight β strands, where the receptor binding site is at one end of the barrel. During the membrane fusion process of influenza virus infection the hemagglutinin molecule undergoes a major structural change in which the fusion peptide moves about 100 Å to a position close to the receptor binding site.

The second protein in the membrane of influenza virus, neuraminidase, does not belong to any of these three groups of barrel structures. Instead, it forms a propeller-like structure of 24 β strands, arranged in six similar motifs that form the six blades of the propeller. Each motif is a β sheet of 4 up-and-down-connected β strands. The enzyme active site is formed by loop regions on one side of the propeller.

In addition to the antiparallel β-structures, there is a novel fold called the β helix. In the β-helix structures the polypeptide chain is folded into a wide helix with two or three β strands for each turn. The β strands align to form either two or three parallel β sheets with a core between the sheets completely filled with side chains.

Petitioner Merck, Ex. 1014, p. 86

Selected readings

General

Chothia, C. Conformation of twisted β-pleated sheets in proteins. *J. Mol. Biol.* 75: 295–302, 1973.

Chothia, C., Janin, J. Orthogonal packing of β-pleated sheets in proteins. *Biochemistry* 21: 3955–3965, 1982.

Chothia, C., Janin, J. Relative orientation of close-packed β-pleated sheets in proteins. *Proc. Natl. Acad. Sci. USA* 78: 4146–4150, 1981.

Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis of the tertiary structure of protein β-sheet sandwiches. *J. Mol. Biol.* 148: 253–272, 1981.

Edison, A.S. Propagation of an error: β-sheet structures. *Trends Biochem. Sci.* 15: 216–217, 1990.

Efimov, A.E. Favoured structural motifs in globular proteins. *Structure* 2: 999–1002, 1995.

Gilbert, W. Why genes in pieces? *Nature* 271: 501, 1978.

Godovac-Zimmerman, J. The structural motif of β-lactoglobulin and retinol-binding protein: a basic framework for binding and transport of small hydrophobic molecules? *Trends Biochem. Sci.* 13: 64–66, 1988.

Jurnak, F., et al. Parallel β domains: a new fold in protein structures. *Curr. Opin. Struct. Biol.* 4: 802–806, 1994.

Lifson, S., Sander, C. Antiparallel and parallel β-strands differ in amino acid residue preference. *Nature* 282: 109–111, 1979.

Lifson, S., Sander, C. Specific recognition in the tertiary structure of β-sheets of proteins. *J. Mol. Biol.* 139: 627–639, 1980.

Ptitsyn, O.B., Finkelstein, A.V. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys.* 13: 339–386, 1980.

Ptitsyn, O.B., Finkelstein, A.V., Falk, P. Principal folding pathway and topology of all-β proteins. *FEBS Lett.* 101: 1–5, 1979.

Richardson, J.S. β-sheet topology and the relatedness of proteins. *Nature* 268: 495–500, 1977.

Richardson, J.S. Handedness of crossover connections in β-sheets. *Proc. Natl. Acad. Sci. USA* 72: 1349–1353, 1975.

Richardson, J.S., Getzoff, E.D., Richardson, D.C. The β-bulge: a common small unit of nonrepetitive protein structure. *Proc. Natl. Acad. Sci. USA* 75: 2574–2578, 1978.

Sawyer, L. Protein structure. One fold among many. *Nature* 327: 659, 1987.

Sibanda, B.L., Blundell, T.L., Thornton, J. Conformation of β-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. *J. Mol. Biol.* 206: 759–777, 1989.

Wilmot, C.M., Thornton, J.M. Analysis and prediction of the different types of β-turns in proteins. *J. Mol. Biol.* 203: 221–232, 1988.

Specific structures

Boumann, U., et al. Three-dimensional structure of the alkaline protease of *Pseudomonas aeruginosa,* a two-domain protein with a calcium binding parallel beta roll motif. *EMBO J.* 12: 3357–3364, 1993.

Bullough, P.A., et al. Structure of influenza haemagglutinin at the pH of membrane fusion. *Nature* 371: 37–43, 1994.

Colman, P.M., Varghese, J.N., Laver, W.G. Structure of the catalytic and antigenic sites in influenza virus neuraminidase. *Nature* 303: 41–44, 1983.

Daniels, R.S., et al. Fusion mutants of the influenza virus hemagglutinin glycoprotein. *Cell* 40: 431–439, 1985.

Jones, T.A., et al. The three-dimensional structure of P2 myelin protein. *EMBO J.* 7: 1597–1604, 1988.

McLachlan, A.D. Repeated folding pattern in copper-zinc superoxide dismutase. *Nature* 285: 267–268, 1980.

Newcomer, M.E., et al. The three-dimensional structure of retinol-binding protein. *EMBO J.* 3: 1451–1454, 1984.

Papiz, M.Z., et al. The structure of β-lactoglobulin and its similarity to plasma retinol-binding protein. *Nature* 324: 383–385, 1986.

Richardson, J.S., et al. Similarity of three-dimensional structure between the immunoglobulin domain and the copper, zinc superoxide dismutase subunit. *J. Mol. Biol.* 102: 221–235, 1976.

Sacchettini, J.C., et al. Refined apoprotein structure of rat intestinal fatty acid binding protein produced in *Escherichia coli. Proc. Natl. Acad. Sci. USA* 86: 7736–7740, 1989.

Steinbacher, S., et al. Crystal structure of P22 tailspike protein: interdigitated subunits in a thermostable trimer. *Science* 265: 383–386, 1994.

Summers, L., et al. X-ray studies of the lens specific proteins. The crystallins. *Pept. Protein Rev.* 3: 147–168, 1984.

Tainer, J.A., et al. Determination and analysis of the 2 Å structure of copper, zinc superoxide dismutase. *J. Mol. Biol.* 160: 181–217, 1982.

Varghese, J.N., Laver, W.G., Colman, P.M. Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9 Å resolution. *Nature* 303: 35–40, 1983.

Watowich, S.I., et al. Crystal structures of influenza virus haemagglutinin in complex with high affinity receptor analogs. *Structure* 2: 719–731, 1994.

Weiss, W., et al. Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid. *Nature* 333: 426–431, 1988.

Petitioner Merck, Ex. 1014, p. 87

Wiley, D.C., Skehel, J.J. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. *Annu. Rev. Biochem.* 56: 365–394, 1987.

Wiley, D.C., Wilson, I.A., Skehel, J.J. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in the antigenic variation. *Nature* 289: 373–378, 1981.

Wilson, I.A., Skehel, J.J., Wiley, D.C. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. *Nature* 289: 366–373, 1981.

Wistow, G., et al. X-ray analysis of the eye lens protein gamma-crystallin at 1.9 Å resolution. *J. Mol. Biol.* 170: 175–202, 1983.

Yoder, M.D., et al. New domain motifs: the structure of pectate lyase C, a secreted plant virulence factor. *Science* 260: 1503–1507, 1993.

Petitioner Merck, Ex. 1014, p. 88

Folding and Flexibility

6

A protein, as we have seen, is a polypeptide chain folded into one or more domains, each of which is made up of α helices, β sheets and loops. The process by which a polypeptide chain acquires its correct three-dimensional structure to achieve the biologically active native state is called protein folding. Although some polypeptide chains spontaneously fold into the native state, others require the assistance of enzymes, for example, to catalyze the formation and exchange of disulfide bonds; and many require the assistance of a class of proteins called chaperones. A chaperone binds to a partly folded polypeptide chain and prevents it from making illicit associations with other folded or partly folded proteins, hence the name chaperone. A chaperone also promotes the folding of the polypeptide chain it holds. After a polypeptide has acquired most of its correct secondary structure, with the α-helices and β-sheets formed, it has a looser tertiary structure than the native state and is said to be in the molten globular state. The compaction that is necessary to go from the molten globular state to the final native state occurs spontaneously.

Protein folding generates a particular three-dimensional structure from an essentially linear, one-dimensional structure—a polypeptide chain with a particular sequence of amino acid residues. How to predict the three-dimensional structure of a protein from its amino acid sequence is the major unsolved problem in structural molecular biology. If we had a general solution to the protein folding problem, it would be possible to write a computer program to simulate protein folding and generate the precise three-dimensional structure of any protein from its amino acid sequence. However, a general solution to the folding problem is still not in sight, even though the number of proteins whose three-dimensional structure has been solved experimentally, in other words, the database of known protein structures, is doubling every 2 years.

A protein in its native state is not static. The secondary structural elements of the domains as well as the entire domains continually undergo small movements in space, either fluctuations of individual atoms or collective motions of groups of atoms. Furthermore, the functional activities of many proteins depend upon large conformational changes triggered by ligand binding. In this chapter, after discussing protein folding, we shall examine some examples of functionally important conformational changes of proteins.

Petitioner Merck, Ex. 1014, p. 89

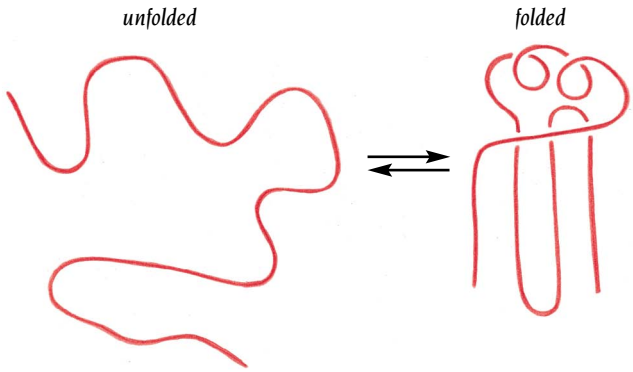


Figure 6.1 A polypeptide chain is extended and flexible in the unfolded, denatured state whereas it is globular and compact in the folded, native state.

Globular proteins are only marginally stable

Every biochemist or molecular biologist who has worked with proteins knows by experience that they are unstable. Slight changes in pH or temperature can convert a solution of biologically active protein molecules in their **native state** to a biologically inactive **denatured state**. The energy difference between these two states in physiological conditions is quite small, about 5–15 kcal/mol, not much more than the energy contribution of a single hydrogen bond, which is of the order of 2–5 kcal/mol.

There are two major contributors to the energy difference between the folded and the denatured state: enthalpy and entropy. **Enthalpy** derives from the energy of the noncovalent interactions within the polypeptide chain—the hydrophobic interactions, hydrogen bonds and ionic bonds. The covalent bonds within and between the amino acid residues in the polypeptide chain are the same in the native and denatured states, with the exceptions of disulfide bonds in those proteins where these form between cysteine residues. The noncovalent interactions on the other hand differ significantly between the two states. In the native state these interactions are maximized to produce a compact globular molecule with a tightly packed hydrophobic core whereas the denatured state is more open and the side chains are more loosely packed (Figure 6.1). These noncovalent interactions are therefore stronger and more frequent in the native state and hence their energy contribution, enthalpy, is much larger. The enthalpy difference between native and denatured states can reach several hundred kcal/mol.

Entropy derives from the second law of thermodynamics which states that energy is required to create order. Proteins in the native state are highly ordered in one main conformation whereas the denatured state is highly disordered, with the protein molecules in many different conformations. A typical experimental preparation of unfolded protein (a solution in 6 M guandinium chloride or 8 M urea) contains 10^{15}–10^{20} protein molecules, each of which will have a unique conformation. In the absence of compensating factors it would therefore be entropically much more favorable for the protein to be in the disordered denatured state. The energy difference due to entropy between the native ordered state and the denatured state can also reach several hundred kcal/mole but in the opposite direction to the enthalpy difference. The total energy difference between the native and the denatured state of 5–15 kcal/mol, which is called the **free energy** difference, is thus a difference between two large numbers, the enthalpy difference and the entropy difference. The fact that this difference is very small is a severe complicating factor both for predictions of possible native states and for interpretation of factors responsible for the stability or instability of protein molecules, because our knowledge about the denatured state is very incomplete.

We know much more about factors that influence the stability of the native state, mainly from experiments using directed mutations in proteins of known three-dimensional structure. Such experiments have yielded

Petitioner Merck, Ex. 1014, p. 90

precise information about energy contributions to the stability of the native state from close packing of hydrophobic side chains in the interior of the protein, and from the presence of disulfide bridges and interior hydrogen bonds and salt bridges, as well as from side chains that compensate the dipole moment of α helices (see Chapter 17).

The marginal stability of the native state over the denatured state is biologically very important. Living cells need globular proteins in correct quantities at appropriate times. It is therefore as important to be able easily to degrade these proteins as it is to be able to synthesize them. Globular proteins in living cells usually have a rather rapid turnover and their native states have therefore evolved to be only marginally stable. Moreover, the catalytic activities of enzymes, and other important functions of proteins, generally require some structural flexibility, which would be inconsistent with a rigidly stabilized structure.

Kinetic factors are important for folding

High resolution x-ray structure determinations of several hundred proteins have shown that in each case the specific sequence of a polypeptide chain appears to yield only a single, compact, biologically active fold in the native state. This fold generally has many substates with minor structural differences between them, as will be discussed later in this chapter, but all of these substates have the same general fold. Comparisons with structure determinations in solution by NMR show that the same fold also prevails in solution. In other words, under physiological conditions there appears to be one conformation for a given amino acid sequence that has a significantly lower free energy than any other. How is this folded state reached?

Intuitively one might imagine that all protein molecules search through all possible conformations in a random fashion until they are frozen at the lowest energy in the conformation of native state. The biophysicist Cyrus Levinthal showed in 1968 by a simple calculation that this is impossible. Assume as a gross simplification that each peptide group has only three possible conformations, the allowed regions α, β and L in the Ramachandran diagram (see Figure 1.7), and that it converts one conformation into another in the shortest possible time, one picosecond (10^{-12} seconds). A polypeptide chain of 150 residues would then have $3^{150} = 10^{68}$ possible conformations. To search all these conformations would require 10^{48} years (10^{56} seconds)—an astronomical number compared with the actual folding time, which is between 0.1 and 1000 seconds both *in vivo* and *in vitro*. To occur on this short time scale, the folding process must be directed in some way through a kinetic pathway of unstable intermediates to escape sampling a large number of irrelevant conformations.

Such a folding mechanism raises several important questions that are difficult to examine experimentally, since the possible intermediates have a very short lifetime. If kinetic factors are important for the folding process it is possible that the observed folded conformation is not the one with the lowest free energy but rather the most stable of those conformations that are kinetically accessible. The protein might be kinetically trapped in a local low energy state with a high energy barrier that prevents it from reaching the global energy minimum which might have a different fold. In such a case structure prediction by energy calculations would give the wrong structure even if such calculations could be made with great accuracy. One important question therefore is how a living cell can prevent the folding pathway from becoming blocked at an intermediate stage. The most common obstacles to correct folding seem to be (1) aggregation of the intermediates through exposed hydrophobic groups, (2) formation of incorrect disulfide bonds, and (3) isomerization of proline residues. To circumvent these three obstacles cells produce special proteins that assist the folding process, as we shall discuss later in this chapter.

Petitioner Merck, Ex. 1014, p. 91

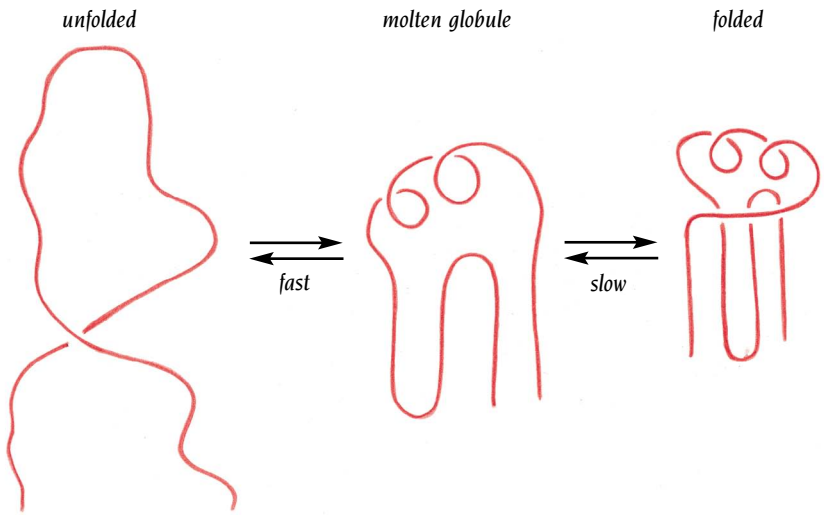


Figure 6.2 The molten globule state is an important intermediate in the folding pathway when a polypeptide chain converts from an unfolded to a folded state. The molten globule has most of the secondary structure of the native state but it is less compact and the proper packing interactions in the interior of the protein have not been formed.

An alternative way to remove kinetic barriers is exemplified by α-lytic protease, a bacterial enzyme which belongs to the serine protease superfamily of enzymes (Chapter 11). Like many other proteases it is synthesized and folded *in vivo* as an inactive precursor protein with a prosegment of 77 residues. This segment is excised after folding to produce the active enzyme. Unfolded precursor protein refolds easily *in vitro* but unfolded α-lytic protease lacking the prosegment does not refold. However, a solution of unfolded enzyme can be induced to refold by adding the excised prosegment. The capacity for folding obviously exists in the unfolded enzyme but there is a barrier present somewhere in the folding pathway that prevents folding. The prosegment removes this kinetic barrier, presumably by interacting with the enzyme in the unfolded state and thereby lowering the free energy of the transition states for folding; just as enzymes lower the free energy of transition states for chemical reactions and thereby increase the rates of the reactions (see Chapter 11).

Molten globules are intermediates in folding

The first observable event in the folding pathway of at least some proteins is a collapse of the flexible disordered unfolded polypeptide chain into a partly organized globular state, which is called the **molten globule** (Figure 6.2). This event is fast, usually within the deadtime of the experimental observation, which is a few milliseconds. We therefore know almost nothing about the process that leads to the molten globule, but we know some of the properties of this state. The molten globule has most of the secondary structure of the native state and in some cases even native-like positions of the α helices and β strands. It is less compact than the native structure and the proper packing interactions in the interior of the protein have not been formed. The interior side chains may be mobile, more closely resembling a liquid than the solid-like interior of the native state. Also loops and other elements of surface structure remain largely unfolded, with different conformations. The molten globule should, therefore, not be viewed as a single structural entity but as an ensemble of related structures that are rapidly interconverting (see Figure 6.3a).

In a second step, which can last up to 1 second, persistent native-like elements of tertiary structure begin to develop, possibly in the form of subdomains that are not yet properly docked. The ensemble of conformations is much reduced compared with those of the molten globule but it is still far from a single form. The single native form is reached in the final stage of folding, which involves the formation of native interactions throughout the protein, including hydrophobic packing in the interior as well as the fixation of surface loops.

Petitioner Merck, Ex. 1014, p. 92

Burying hydrophobic side chains is a key event

The collapse of the unfolded state to generate the molten globule embodies the main mystery of protein folding. What is the driving force behind the choice of native tertiary fold from a randomly oriented polypeptide chain?

There is very little change in free energy by forming the internal hydrogen bonds that are characteristic of α helices and β sheets because in the unfolded state equally stable hydrogen bonds can be formed to water molecules. Secondary structure formation therefore cannot be the thermodynamic driving force of protein folding. On the other hand there is a large free energy change by bringing hydrophobic side chains out of contact with water and into contact with each other in the interior of a globular entity. Thus the most likely scenario is that the polypeptide chain begins to form a compact shape with hydrophobic side chains at least partially buried very early in the folding process. This scenario has several important consequences. It vastly reduces the number of possible conformations that need to be searched because only those that are sterically accessible within this shape can be sampled. Second, when some of the side chains are partly buried, their polar backbone -NH and -CO groups are also buried in a hydrophobic environment unable to form hydrogen bonds to water. This is energetically unfavorable unless they form hydrogen bonds to each other, which they can only do if they are close together. The simplest way to form such bonds is by forming elements of secondary structure: α helices and β sheets. The formation of secondary structure early in the folding process can therefore be regarded as a consequence of burying hydrophobic side chains and not as a driving force for the formation of the molten globule.

Looking at the amino acid sequence of a globular protein one finds that hydrophobic side chains are usually scattered along the entire sequence in a seemingly random manner. In the native state of the folded protein about half of these side chains are buried in the interior and the rest are scattered on the surface of the protein, surrounded by hydrophilic side chains. The buried hydrophobic side chains are not clustered in the sequence but are scattered along the entire polypeptide chain. What causes these residues to be selectively buried during the early and rapid formation of the molten globule? This question must be answered before one can solve the folding problem and be able to predict the fold of a protein from its amino acid sequence.

Both single and multiple folding pathways have been observed

In order to understand fully any folding pathway, all states of the pathway must be characterized both structurally and energetically. The simplified diagram in Figure 6.3 illustrates that during the folding process the protein proceeds from a high energy unfolded state to a low energy native state through metastable intermediate states with local low energy minima separated by unstable transition states of higher energy. The characterization of these states is not trivial and many different experimental techniques are employed, including NMR, hydrogen exchange, spectroscopy and thermochemistry.

Recently Alan Fersht, Cambridge University, has developed a protein engineering procedure for such studies. The technique is based on investigation of the effects on the energetics of folding of single-site mutations in a protein of known structure. For example, if minimal mutations such as Ala to Gly in the solvent-exposed face of an α helix, destabilize both an intermediate state and the native state, as well as the transition state between them, it is likely that the helix is already fully formed in the intermediate state. If on the other hand the mutations destabilize the native state but do not affect the energy of the intermediate or transition states at all, it is likely that the helix is not formed until after the transition state.

Petitioner Merck, Ex. 1014, p. 93

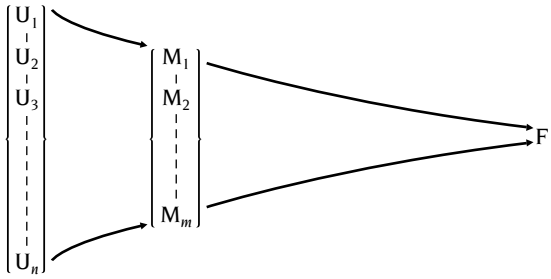
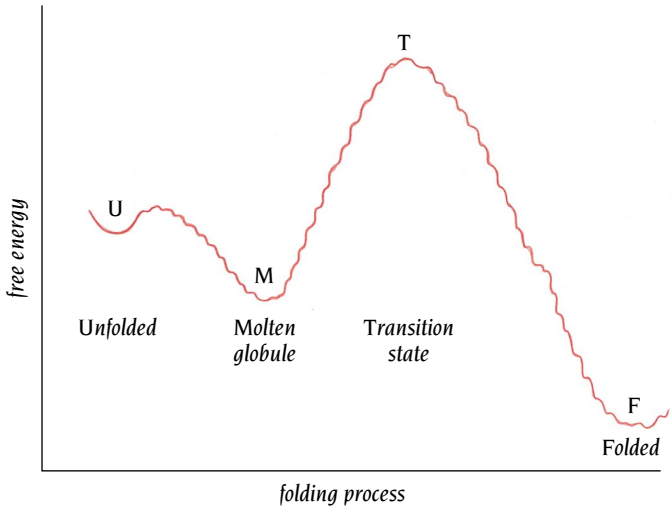


Figure 6.3 The unfolded state is an ensemble of a large number of conformationally different molecules, $U_1 \ldots U_n$, which undergo rapid interconversions. The molten globule is an ensemble of structurally related molecules, $M_1 \ldots M_m$, which are rapidly interconverting and which slowly change to a single unique conformation, the folded state F. During the folding process the protein proceeds from a high energy unfolded state to a low energy native state. The conversion from the molten globule state to the folded state is slow and passes through a high energy transition state, T.



The small bacterial ribonuclease, **barnase**, is a single chain protein with 110 amino acids and no disulfide bridges. Its three-dimensional structure was determined by the group of Guy Dodson, York University, and comprises three amino terminal α helices and a carboxy terminal five-stranded antiparallel β sheet (Figure 6.4). The group of Alan Fersht have examined the effects of mutations all along the structure and have made a detailed residue by residue characterization of its folding intermediate and transition states. They have concluded from their results that the intermediate molten globule state already has not only most of the native secondary structure elements but also the native-like relative positions of the α helix and β sheet as well as

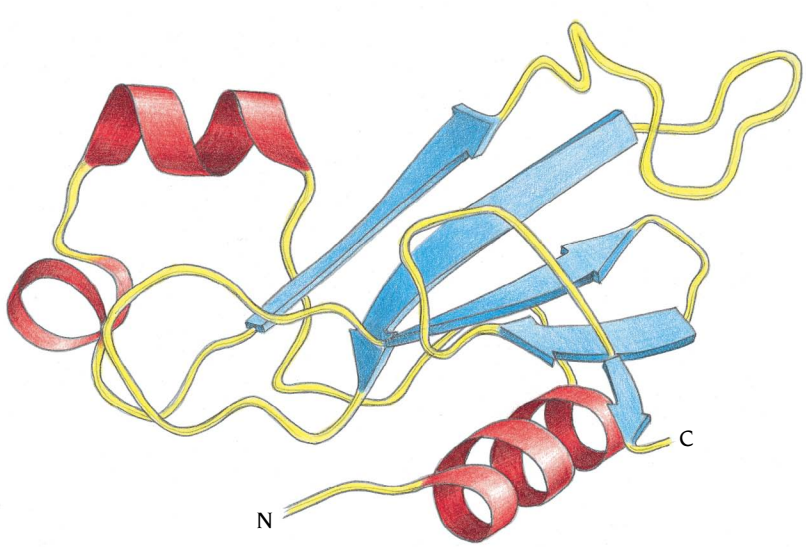


Figure 6.4 Schematic diagram of the structure of the enzyme barnase which is folded into a five stranded antiparallel β sheet (blue) and two α helices (red).

Petitioner Merck, Ex. 1014, p. 94

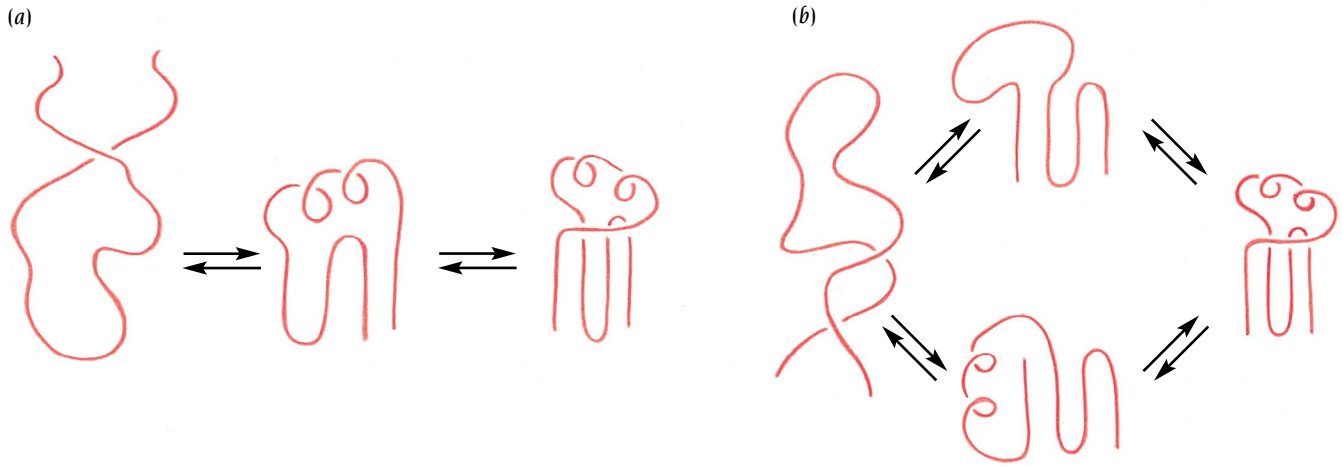


Figure 6.5 (a) Some proteins such as barnase fold through one major pathway whereas others fold through multiple pathways. (b) The folding of the enzyme lysozyme proceeds through at least two different pathways.

the relative positions of the β strands within the sheet. These results are consistent with the notion that the folding of barnase proceeds through a single major transition state and consequently through one major pathway (Figure 6.5a).

In contrast, folding of the enzyme **lysozyme** involves parallel pathways and distinct folding domains. Hen egg-white lysozyme was the first enzyme to have its structure determined crystallographically, in the laboratory of David Phillips then at the Royal Institution, London in 1965. The native structure consists of two lobes separated by a cleft (Figure 6.6). The first lobe comprises five α helices and the second is predominantly a three-stranded antiparallel β sheet. The folding of lysozyme has been studied extensively by a variety of complementary techniques (NMR, circular dichroism, fluorescence, hydrogen–deuterium exchange) to follow the development of different aspects of the structure such as formation of secondary structure, burial of hydrophobic aromatic groups and formation of hydrogen bonds. The group of Christopher Dobson, Oxford University, has used pulsed amide hydrogen–deuterium exchange to follow secondary structure formation. Amide hydrogen atoms are readily exchanged with the solvent in unfolded proteins, but this exchange is often strongly inhibited in a folded protein, especially for those amide groups that are hydrogen bonded in secondary structure elements. As a result, by measuring the rate of amide–hydrogen exchange as a function of folding time it is possible to monitor the formation of structure during the folding reaction. At 20 milliseconds, two major intermediate stages of lysozyme were detected: one in which the α-helical domain

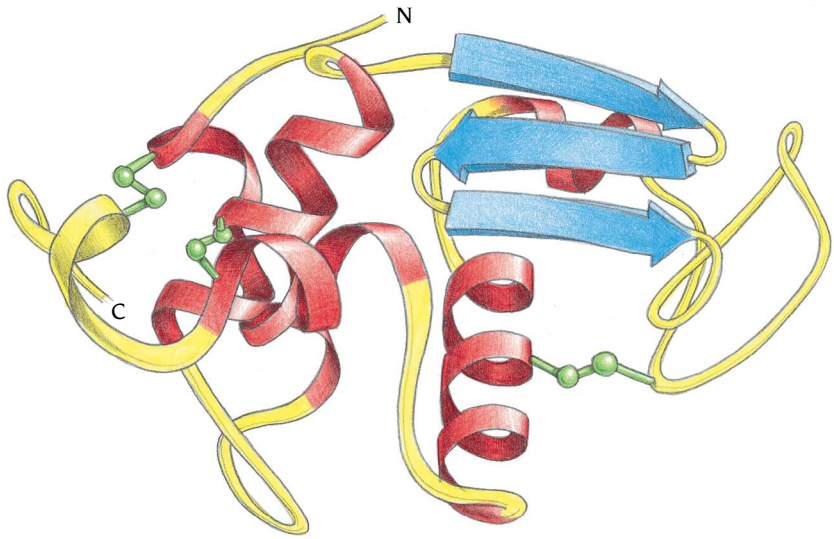


Figure 6.6 Schematic diagram of the structure of the enzyme lysozyme which folds into two domains. One domain is essentially α-helical whereas the second domain comprises a three stranded antiparallel β sheet and two α helices. There are three disulfide bonds (green), two in the α-helical domain and one in the second domain.

Petitioner Merck, Ex. 1014, p. 95

had achieved a high degree of secondary structure while the β-sheet domain contained no detectable structure, and a second state in which no stable structure was observed in either the α or the β domain. In addition, a third, less populated state was observed in which both the α domain and the β domain had significant structure. Therefore in this case it is likely that the folding follows different pathways, two major and one minor, with different molten globule states which in a later stage of the folding process converge to one final native state (see Figure 6.5b).

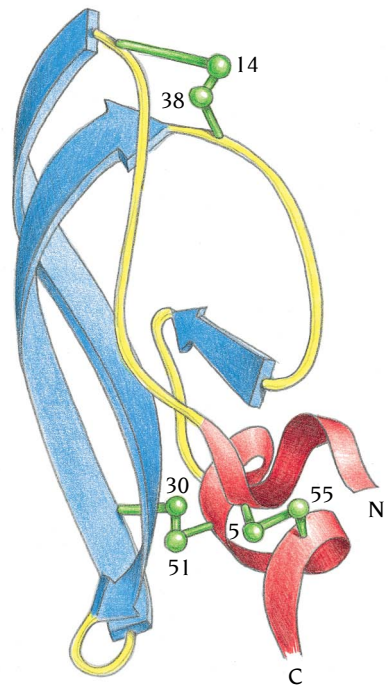
Enzymes assist formation of proper disulfide bonds during folding

The formation of correct disulfide bonds during the folding process poses special problems for cells. In the denatured, unfolded state of proteins there are no disulfide bridges; the cysteine residues are reduced. Formation of disulfide bonds requires oxidation of the cysteine residues. In bacteria this oxidation occurs mainly in the periplasmic space and is catalyzed by a family of enzymes called disulfide bridge-forming enzymes, **Dsb**. In bacteria proteins with disulfide bridges are therefore essentially only found in the periplasmic space and in the outer membrane or they are secreted. In eucaryotic cells disulfide bond formation occurs in the endoplasmic recticulum before proteins are exported to the cell surface. Here an enzyme called **protein disulfide isomerase**, **PDI**, catalyzes internal disulfide exchange to remove folding intermediates with incorrectly formed disulfide bridges. Again, proteins with disulfide bonds are not found in the cytosol but are located in the plasma membrane or are secreted.

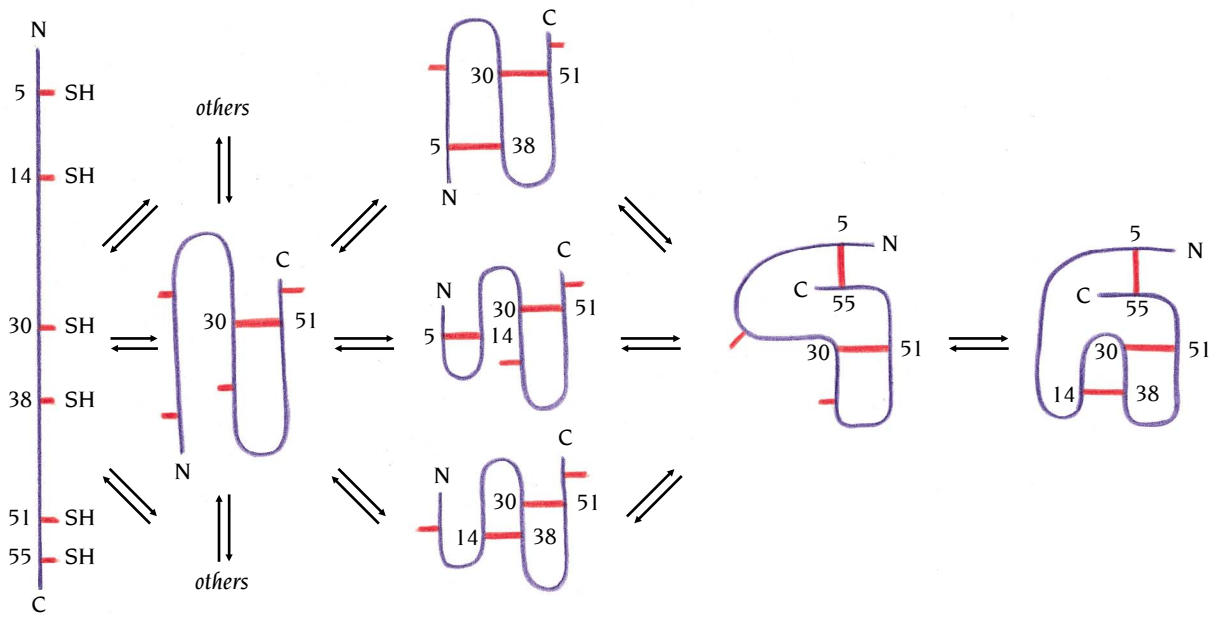
The influence of disulfide bond formation on folding *in vitro* has been extensively studied by Thomas Creighton, EMBL, Heidelberg, and more recently by Peter Kim at MIT. Creighton's pioneering work, which introduced the trapping of disulfide-bonded intermediates as a method for studying the folding pathways of proteins, as well as recent experiments, have shown that these *in vitro* results are also relevant for the folding process *in vivo*. Both Creighton and Kim have studied a small eucaryotic protein, **bovine pancreatic trypsin inhibitor**, **BPTI**, which has six cysteine residues that form three disulfide bonds within its polypeptide chain of 58 residues (Figure 6.7a).

Figure 6.7 (a) Schematic diagram of the structure of the small protein bovine pancreatic trypsin inhibitor, BPTI. The three disulfide bonds are green, β strands are blue and α helices are red. (b) The folding pathway of BPTI according to Creighton. The unfolded protein has six cysteine residues in their reduced state. The major single disulfide bond intermediate has a disulfide bond between residues 30 and 51. This intermediate forms double disulfide bond intermediates which contain non-native disulfide bonds. According to Creighton these intermediates are essential for the formation of the native double disulfide bond intermediate 30–51, 5–55 which rapidly forms the third native disulfide bond 14–38.

(*a*)



(*b*)



Petitioner Merck, Ex. 1014, p. 96

The fully reduced protein is largely unfolded and does not fold until the cysteine residues are oxidized to disulfide bridges. In the native state these bonds are between cysteine residues 30–51, 5–55 and 14–38. During the folding process formation of the first disulfide bond is almost random and the 15 possible single-disulfide species are in rapidly interchanging equilibrium. However, the intermediate with the disulfide bond at 30–51 is more stable than the other 14 and is present in about 60% of the molecules. The productive folding pathway goes through this stable intermediate, consequently all other intermediates must eventually rearrange their disulfide bond in order to fold properly. It has a partly folded conformation comprising the native-like α helix and β sheet linked by the 30–51 disulfide bond. The remaining part of the polypeptide chain is disordered or unfolded and includes cysteines 5, 14 and 38 but not cysteine 55 which is within the folded α helix.

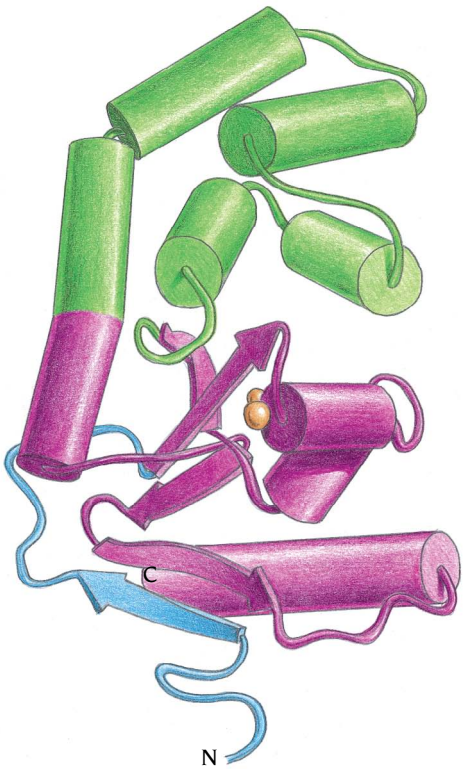
The second disulfide bonds formed in the 30–51 intermediate are between all three possible pairs of flexible cysteines, to form 5–14, 5–38 or 14–38 disulfide bonds (Figure 6.7b). The first two are non-native disulfide bonds, which are not stabilized to any significant extent, and occur primarily because they are in flexible parts of the polypeptide chain. In contrast, upon forming the 14–38 disulfide bond, there are now two native disulfide bonds, and the molecule adopts a very native-like, folded conformation. However, formation of the last disulfide bond between cysteines 5 and 55 occurs only very slowly in this intermediate because these cysteine residues are buried inside the folded intermediate and not accessible to oxidizing agents or disulfide rearrangement. The productive precursor to the native state is instead the intermediate with the 30–51 and 5–55 disulfide bonds since the remaining cysteines 14 and 38 are in good proximity on the surface of the folded protein and there is no barrier to them forming the final disulfide bond. This intermediate can be reached either by rearrangement of the non-native disulfide bonds formed in the 30–51 intermediate or by unfolding the 30–51, 14–38 intermediate.

The folding pathway of BPTI illustrates clearly the importance of disulfide rearrangements for the folding process and hence the advantage for a cell of having enzymes that increase the rate of this reaction. Adding the enzyme protein disulfide isomerase significantly increases the rate of folding of BPTI *in vitro*. The three-dimensional structure of a eucaryotic protein disulfide isomerase has not yet been determined but the structure of DsbA from *Escherichia coli*, a member of the Dsb family of enzymes, has been solved by John Kuriyan, Rockefeller University. This monomeric enzyme has 189 amino acid residues that fold into two domains; one comprises five tightly packed α helices while the second domain has a structure very similar to that of **thioredoxin** (Figure 6.8). Thioredoxin, the x-ray structure of which was determined by the group of Carl Branden, Uppsala, is a ubiquitous protein that functions as a general protein disulfide oxido-reductase and in bacteria is responsible for keeping disulfide bridges reduced. The well-characterized mechanism of thioredoxin is based on reversible oxidation of two cysteine thiol groups to a disulfide, accompanied by the transfer of two electrons and two protons. Presumably DsbA functions in a similar way since the redox-active disulfide bridges in these two proteins are very similar.

The thioredoxin domain (see Figure 2.7) has a central β sheet surrounded by α helices. The active part of the molecule is a βαβ unit comprising β strands 2 and 3 joined by α helix 2. The redox-active disulfide bridge is at the amino end of this α helix and is formed by a Cys-X-X-Cys motif where X is any residue in DsbA, in thioredoxin, and in other members of this family of redox-active proteins. The α-helical domain of DsbA is positioned so that this disulfide bridge is at the center of a relatively extensive hydrophobic protein surface. Since disulfide bonds in proteins are usually buried in a hydrophobic environment, this hydrophobic surface in DsbA could provide an interaction area for exposed hydrophobic patches on partially folded protein substrates.

Disulfide bonds in proteins are generally stable and nonreactive, acting like bolts in the structure. However, oxidized DsbA is less stable than the reduced form and its disulfide bond is very reactive. DsbA is thus a strong

Figure 6.8 Schematic diagram of the enzyme DsbA which catalyzes disulfide bond formation and rearrangement. The enzyme is folded into two domains, one domain comprising five α helices (green) and a second domain which has a structure similar to the disulfide-containing redox protein thioredoxin (violet). The N-terminal extension (blue) is not present in thioredoxin. (Adapted from J.L. Martin et al., *Nature* 365: 464–468, 1993.)



Petitioner Merck, Ex. 1014, p. 97

oxidizing agent which is consistent with its ability to catalyze the exchange of disulfide bonds in other proteins. Although the structure of eucaryotic protein disulfide isomerase is unknown, sequence alignment shows that its 189 amino acid polypeptide chain contains two domains that are homologous to thioredoxin. This indicates that both the bacterial and the eucaryotic proteins that catalyze the formation of disulfide bonds in proteins during folding use thioredoxin-like domains and presumably are evolutionarily related.

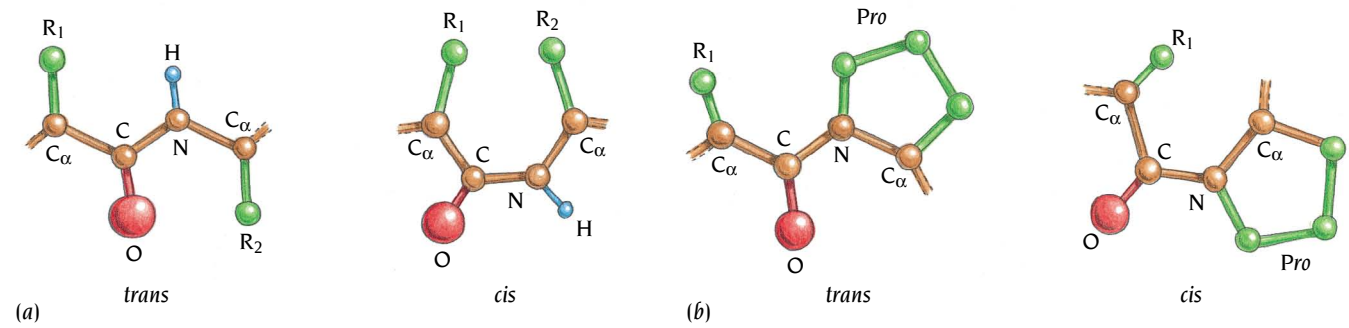
Isomerization of proline residues can be a rate-limiting step in protein folding

In Chapter 1 we presented a picture of the peptide as a planar unit with the C=O and N-H groups pointing in opposite directions in the plane (see Figure 1.5). This so called *trans*-peptide is the most stable form of the peptide group. However, there is another possible form, the *cis*-peptide, which is also planar but in which the C=O and N-H groups point in the same direction (Figure 6.9a). For most peptides the *cis*-form is about 1000 times less stable than the *trans*-form and consequently *cis*-peptides are rarely found in native proteins. However, when the second residue (R_2 in Figure 6.9a) is a proline the *cis*-form is only about four times less stable than the *trans*-form (Figure 6.9b) and some *cis*-proline peptides occur in many proteins. Most proline residues in proteins are of course in the *trans*-configuration because in *trans* there are fewer steric collisions, but *cis*-prolines are found in tight bends of the polypeptide chain and are sometimes essential for activity or for conformational flexibility.

In the native protein these less stable *cis*-proline peptides are stabilized by the tertiary structure but in the unfolded state these constraints are relaxed and there is an equilibrium between *cis*- and *trans*-isomers at each peptide bond. When the protein is refolded a substantial fraction of the molecules have one or more proline-peptide bonds in the incorrect form and the greater the number of proline residues the greater the fraction of such molecules. *Cis–trans* isomerization of proline peptides is intrinsically a slow process and *in vitro* it is frequently the rate-limiting step in folding for those molecules that have been trapped in a folding intermediate with the wrong isomer.

In vivo the rate of this process is enhanced by enzymes initially called **peptidyl prolyl isomerases**, which are found in both procaryotic and eucaryotic organisms. Surprisingly these enzymes were later found to be involved in immunosuppression by inhibiting T cell proliferation after binding of immunosuppressive drugs; this medically important activity of peptidyl prolyl isomerases is unrelated to their isomerase activity in folding. The first peptidyl prolyl isomerase to be discovered is now called **cyclophilin** because of its role as target for the most frequently used therapeutic agent for prevention of graft rejection, cyclosporin A. Cyclophilin is an abundant soluble protein consisting of a single polypeptide chain of 165 amino acids. Because it is a target for cyclosporin A the determination of its structure was pursued in the early 1990s by several pharmaceutical companies intent

Figure 6.9 (a) Peptide units can adopt two different conformations, *trans* and *cis*. In the *trans*-form the C=O and the N-H groups point in opposite directions whereas in the *cis*-form they point in the same direction. For most peptides the *trans*-form is about 1000 times more stable than the *cis*-form. (b) When the second residue in a peptide is proline the *trans*-form is only about four times more stable than the *cis*-form. *Cis*-proline peptides are found in many proteins.



Petitioner Merck, Ex. 1014, p. 98

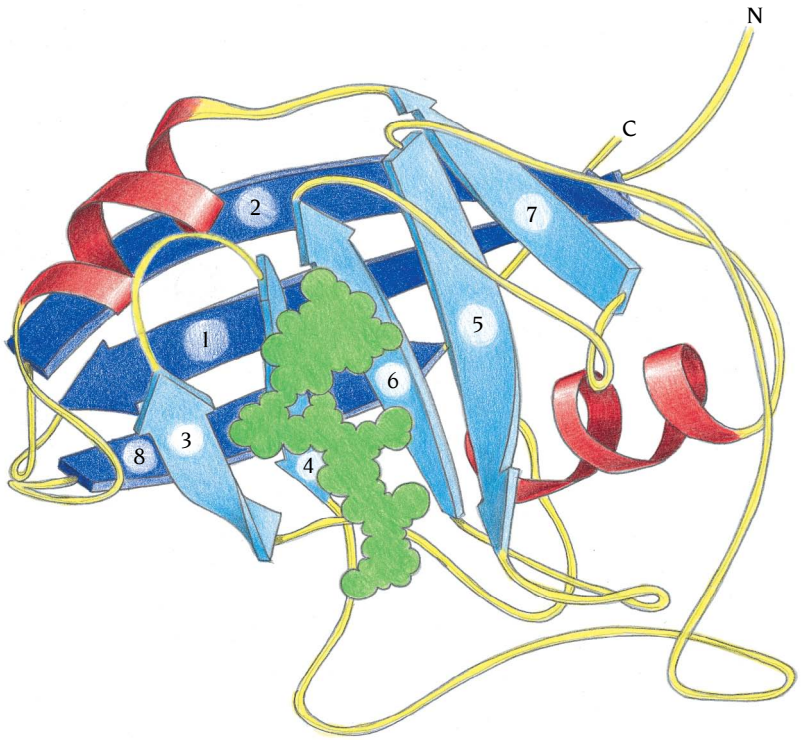


Figure 6.10 Schematic diagram of the structure of the protein cyclophilin, a prolyl peptide isomerase that catalyzes the conversion between *cis*- and *trans*-proline peptides. The protein is folded into an eight-stranded antiparallel β barrel (blue) with two α helices (red) on the outside. The active site, which has been located by the binding of a proline containing tetrapeptide (green), is on the outside of the β barrel. (Adapted from J. Kallen et al., *Nature* 353: 276–279, 1991.)

upon designing better immunosuppressive drugs, and a group from Sandoz, Basel, in collaboration with Kurt Würthrich, ETH, Zurich, published the first structure of cyclophilin with a bound proline-containing peptide using a combination of x-ray crystallography and NMR.

Their work showed that the structure of cyclophilin complexed with a proline-containing tetrapeptide (Figure 6.10) is essentially an eight-stranded β barrel with an unusual topology, different from both the up-and-down and the Greek key β barrels discussed in Chapter 5. The interior of the barrel is tightly packed with hydrophobic residues. In contrast to the up-and-down β barrels that bind ligands inside the barrel, the tetrapeptide ligand of cyclophilin is bound on the outside of the barrel in a deep groove between one face of the β sheet and a long loop region (see Figure 6.10).

Cyclophilin enhances the rate of *cis–trans* isomerization of proline peptides by a factor of one million over the nonenzymatic rate. The detailed mechanism of the enzymatic process is still unclear but mechanisms based on distortion or desolvation of the peptide group are possible. The proline residue is tightly bound in a hydrophobic pocket of the binding groove and the carbonyl oxygen atom is hydrogen-bonded to basic side chains. The binding of a peptide segment into a hydrophobic environment may promote *cis–trans* isomerization by decreasing the charge separation in the peptide group (see Figure 2.3a) and thus creating a peptide bond with a more single-bond character. Alternatively, the tight hydrophobic interactions with cyclophilin might distort the geometry of the peptide group toward the transition state for isomerization. Both mechanisms or a combination of them will decrease the energy barrier for rotation around the peptide bond, which is required for the *cis–trans* isomerization.

Proteins can fold or unfold inside chaperonins

Before protein molecules attain their native folded state they may expose hydrophobic patches to the solvent. Isolated purified proteins will aggregate during folding even at relatively low protein concentrations. Inside cells, where there are high concentrations of many different proteins, aggregation could therefore occur during the folding process. This is prevented by

Petitioner Merck, Ex. 1014, p. 99

molecular chaperones, ubiquitous and abundant families of proteins that assist the folding of both nascent polypeptides still attached to ribosomes and released completed polypeptide chains. Several chaperones are induced by heat shock, because protein unfolding and aggregation is increased at elevated temperatures; and consequently chaperones were first classified as **heat-shock proteins** (Hsps). Several of these heat-shock proteins, such as the 70 kDa protein **DnaK** (**Hsp 70**) are also expressed normally and participate in various other cellular processes including assembly and disassembly of multimeric proteins and membrane translocation of secreted proteins. The Hsp 70 polypeptide chain is divided into two functional regions, one that binds and hydrolyses ATP and a second that binds hydrophobic segments of unfolded polypeptide chains. The N-terminal ATP-binding region has a four-domain structure very similar to that of actin (Chapter 14); the polypeptide-binding domain is an antiparallel β sandwich in the C-terminal region. There is as yet no structural information on the complete Hsp 70 molecule. By contrast, the role during folding of two other heat-shock proteins, Hsp 60 and Hsp 10, with molecular weights of 60 kDa and 10 kDa, respectively, has been studied extensively, especially in *Escherichia coli* where they are called **GroEL** and **GroES**, respectively. These proteins function together as a large complex, called a **chaperonin**, consisting of 14 subunits of GroEL and 7 subunits of GroES and requiring ATP to function.

Chaperonins bind unfolded, partly folded and incorrectly folded protein molecules but not proteins in their native state. They are promiscuous in that they bind to and assist the folding of a large number of different proteins independent of the latter's amino acid sequences. How can these chaperonins distinguish between correctly and incorrectly folded versions of almost any water-soluble polypeptide chain and how can they mediate the efficient conversion of unfolded or misfolded proteins to their native form? The x-ray structure determinations of GroEL and GroES, in combination with electron microscopic studies of GroEL–GroES–polypeptide complexes, have made possible major steps towards understanding these processes.

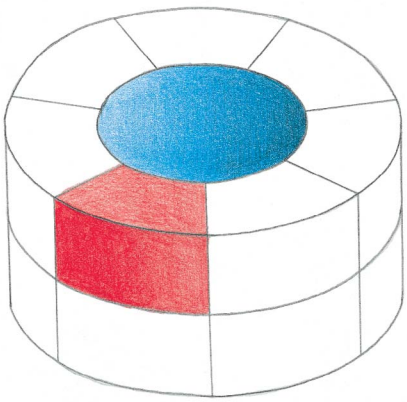


Figure 6.11 Schematic diagram of the chaperonin GroEL molecule as a cylinder with 14 subunits arranged in two rings of 7 subunits each. The space occupied by one subunit is red and the hole inside the cylinder is blue.

GroEL is a cylindrical structure with a central channel in which newly synthesized polypeptides bind

The x-ray structure of GroEL was determined in 1994 by the groups of Paul Sigler and Arthur Horwich, Yale University. The 14 subunits of GroEL, each comprising 547 amino acids, form two rings in which the 7 subunits of each ring are arranged with nearly exact sevenfold rotational symmetry. The rings are arranged back-to-back, forming an extensive interface with one another across an almost flat equatorial plane. The whole structure resembles a porous thick-walled cylinder that is about 150 Å long and 140 Å in diameter and contains a large central cavity or channel (Figure 6.11). Each subunit has three distinct domains, equatorial, intermediate and apical, that are arranged in the cylinder as shown in Figure 6.12. The equatorial domain which is largest and comprises 243 residues is mainly α-helical. It serves as the foundation of the GroEL structure providing all of the contacts between the two sevenmembered rings across the equatorial plane. In addition it provides most of the contacts between the subunits within each ring as well as the ATP binding site, which is essential for function. The equatorial domain contains both the N-terminus and the C-terminus of the polypeptide chain. About 30 residues at these ends are not visible in the x-ray structure; either they are disordered, or they occupy differently ordered positions in different molecules. The visible ends point into the central cavity close to the equatorial plane and presumably the nonvisible residues partially block the channel in this region; three-dimensional electron microscopic reconstruction of individual chaperonin particles shows that the channel is narrow in the center of the molecule (Figure 6.13).

The apical domain (residues 191–376) is essentially a four-layer structure comprising two β sheets sandwiched between α helices. One β sheet has

Petitioner Merck, Ex. 1014, p. 100

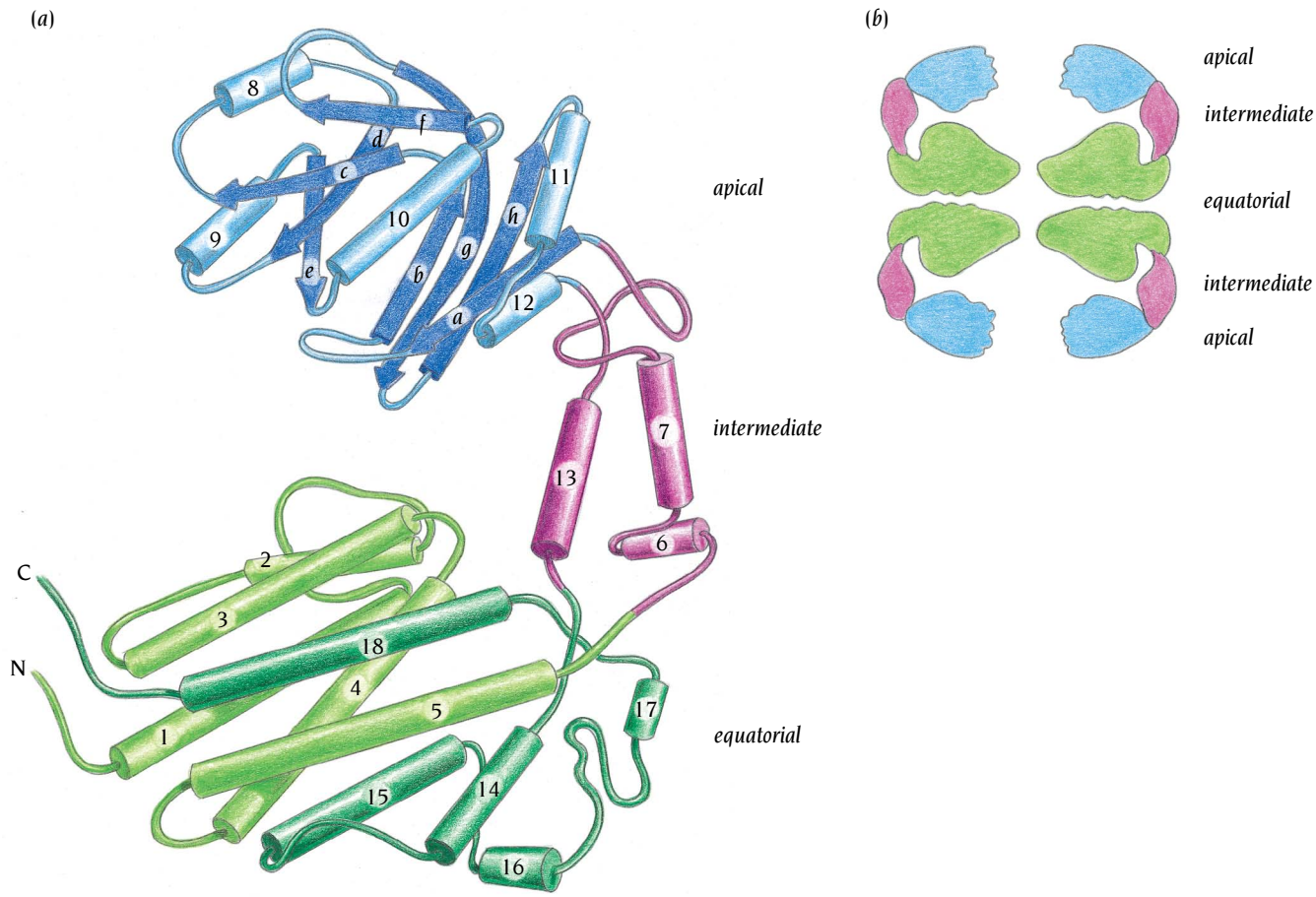


Figure 6.12 (a) Schematic diagram of one subunit of GroEL. The polypeptide chain is folded into three domains. The equatorial domain (green) is the largest domain, comprising 10 α helices, and is built up from both the N-terminal and the C-terminal regions. The apical domain (blue), which is a β sandwich flanked by α helices, is formed by the middle region of the polypeptide chain. The two linker regions between the equatorial and the apical domains form a small intermediate domain (purple) comprising three α helices. (b) Schematic diagram illustrating the domain arrangement of four subunits in the GroEL molecule, two in each of the seven membered rings. The equatorial domains form the middle part of the molecule and interact with each other both within each ring and between the rings. The apical domains are at the top and the bottom of the cylinder and form an opening to the interior of the molecule. The small intermediate domains form the thinnest part of the cylinder wall in the middle of each ring. [(a) Adapted from K. Braig et al., *Nature* 371: 578–586, 1994.]

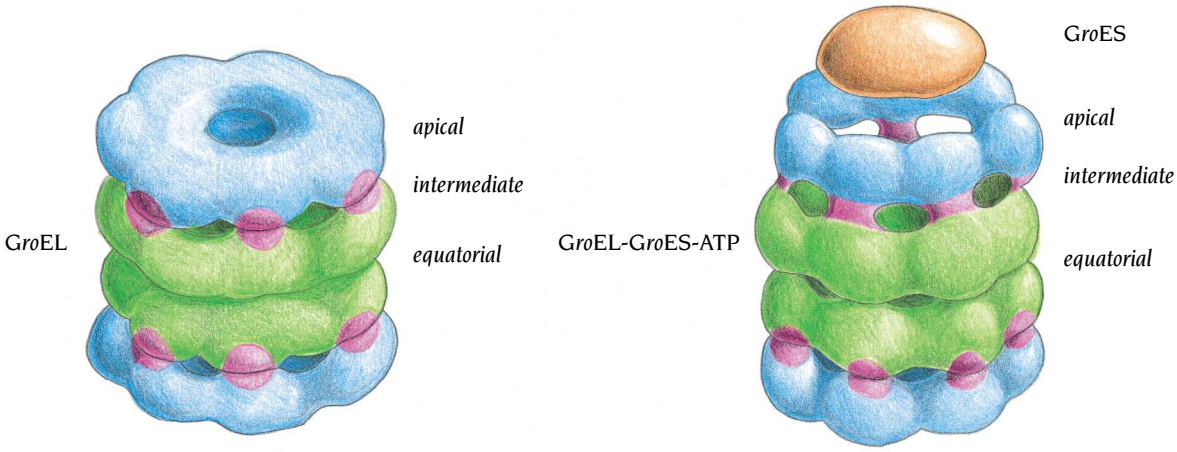


Figure 6.13 Models of the GroEL molecule in two different functional states based on three-dimensional reconstruction from electron microscopy pictures. A large conformational change of GroEL occurs when GroES and ATP are bound. The GroES molecule binds to one of the GroEL rings and closes off the central cavity. The GroEL ring becomes larger and the cavity inside that part of the cylinder becomes wider. (Adapted from S. Chen et al., *Nature* 371: 261–264, 1994.)

Petitioner Merck, Ex. 1014, p. 101

antiparallel β strands whereas the other is part of an α/β domain with four parallel β strands. The apical domains form the opening to the solvent of the central channel. The segments of this domain that form the top surface of the molecule as well as those that face the upper regions of the channel are flexible and not very well ordered. These segments are rich in hydrophobic residues and they are involved in binding to the hydrophobic areas exposed by non-native folds of polypeptide chains. Mutating these hydrophobic residues to charged ones abolishes both GroES and polypeptide binding whereas mutations conserving hydrophobic residues, such as Phe to Val, have no functional effects. The flexibility of these segments probably accounts for the promiscuous binding of a wide range of unfolded polypeptide sequences.

The equatorial and apical domains are linked by a small intermediate domain which forms part of the outer wall and extends only about 25 Å in the radial direction; consequently the internal cavity is wider in this region, up to 90 Å in diameter. In addition there are holes in the wall between the intermediate domains in adjacent subunits. These seven holes are large enough to permit ATP and ADP to diffuse in and out. The intermediate domain is connected to the other two domains through short antiparallel segments that could easily serve as hinges during conformational changes. Electron microscopic studies of GroEL with different ligands bound have shown that substantial changes in the orientations of the domains and in the size of the central cavity occur during the functional cycle of the chaperonin (see Figure 6.13).

GroES closes off one end of the GroEL cylinder

GroES binds to the apical domain of GroEL, closing off the central cavity (see Figure 6.13). Once GroES has bound to one of the rings in the GroEL molecule a conformational change occurs which decreases the affinity of the second GroEL ring for GroES. The predominant functional state of the GroEL–GroES complex is, therefore, asymmetric with GroES bound to only one end of the GroEL cylinder. Obviously there is strong structural communication between the halves of the GroEL molecule since GroES binding to one half affects the properties of the other half, despite a distance of about 150 Å between the two GroES binding sites.

The GroES molecule comprises seven subunits, each of 97 amino acids linked together around a sevenfold rotation axis, the same symmetry arrangement as in one GroEL ring. The x-ray structure of GroES was determined in 1996 by the group of Johann Deisenhofer, University of Texas, Dallas. The GroES molecule is shaped like a dome, about 75 Å in diameter and 30 Å high with a small hole in the middle (Figure 6.14a). The core of the subunit structure is a β barrel comprising two antiparallel β sheets packed against each other (Figure 6.14b). Two large loop regions protrude from this core, one of which extends above the plane of the ring creating a loosely packed top of the dome that covers the central hole. The other loop region, which is rich in hydrophobic residues, extends below the dome and presumably interacts with the apical domain in the GroES–GroEL complex. This loop is disordered in the x-ray structure of GroES but NMR studies have shown that the loop becomes ordered when GroES binds to GroEL and mutational studies have shown that the hydrophobic residues in this loop are important for chaperonin function.

The GroEL–GroES complex binds and releases newly synthesized polypeptides in an ATP-dependent cycle

How does the GroEL–GroES complex function as a chaperone to assist protein folding? Although several aspects of the mechanism are not clear, the main features of the functional cycle are known. The first step is the

Petitioner Merck, Ex. 1014, p. 102

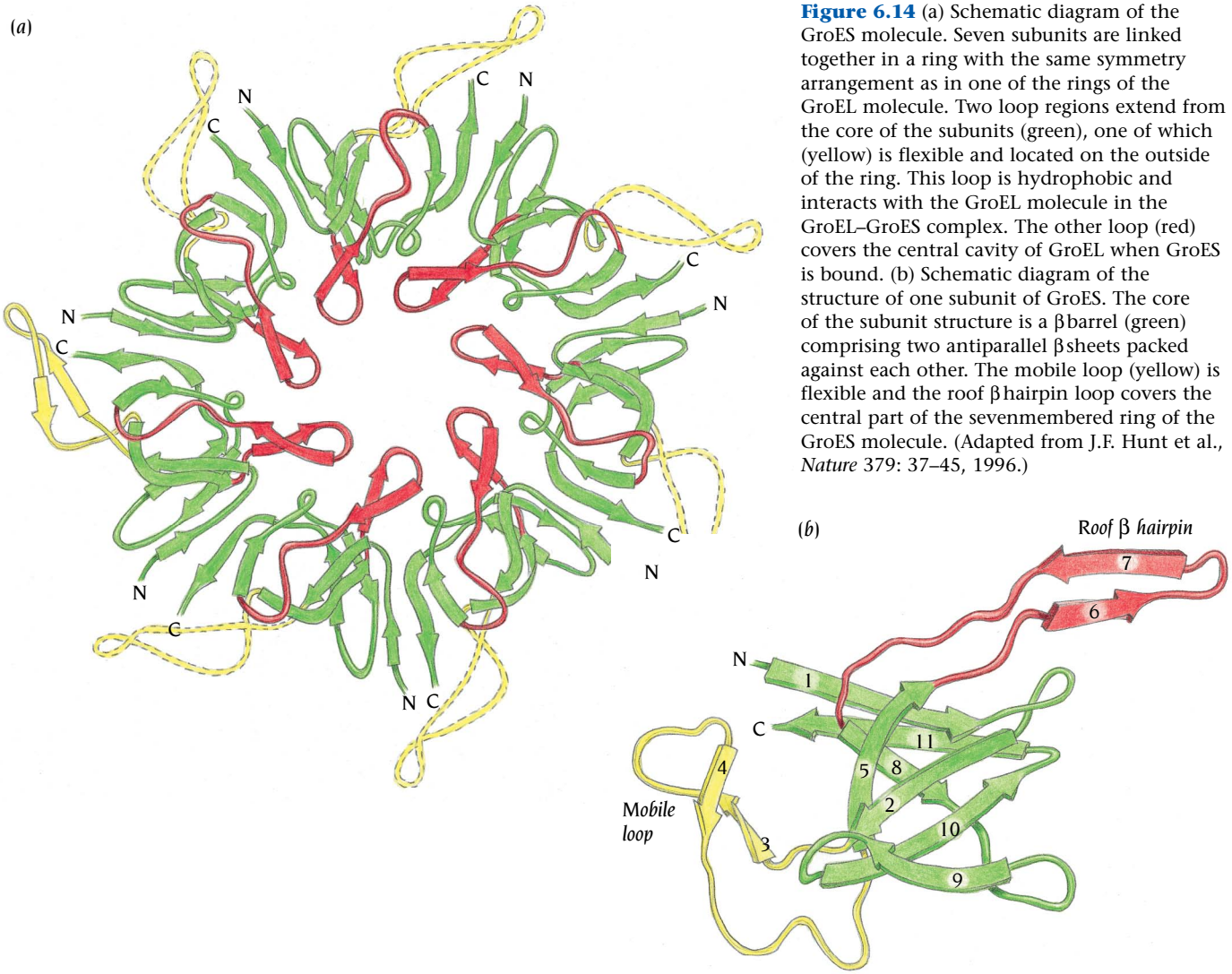


Figure 6.14 (a) Schematic diagram of the GroES molecule. Seven subunits are linked together in a ring with the same symmetry arrangement as in one of the rings of the GroEL molecule. Two loop regions extend from the core of the subunits (green), one of which (yellow) is flexible and located on the outside of the ring. This loop is hydrophobic and interacts with the GroEL molecule in the GroEL–GroES complex. The other loop (red) covers the central cavity of GroEL when GroES is bound. (b) Schematic diagram of the structure of one subunit of GroES. The core of the subunit structure is a β barrel (green) comprising two antiparallel β sheets packed against each other. The mobile loop (yellow) is flexible and the roof β hairpin loop covers the central part of the sevenmembered ring of the GroES molecule. (Adapted from J.F. Hunt et al., *Nature* 379: 37–45, 1996.)

formation of a GroEL–ATP complex, one end of which which then binds one molecule of GroES, with the hydrolysis of ATP. The resulting GroEL–ADP–GroES is a stable complex the halves of which have very different properties (Figure 6.15a). The GroEL ring where GroES is bound (*cis*-position) has undergone a large structural change forming a wide internal cavity whose walls are formed from both the apical and intermediate domains. This cavity is partly closed off from the solvent by the GroES dome. The other ring (*trans*-position) has a smaller cavity which is open to the solvent. Unfolded proteins can bind in both the *cis*- and the *trans*-positions but only those that are bound in the *cis*-position undergo subsequent folding rearrangements (Figure 6.15c–e). Binding and release of polypeptides at the *trans*-position seem to be functionally unimportant.

Release of bound polypeptides from the closed cavity in the *cis*-position requires that GroES is first released. The release of GroES, like its binding, requires ATP hydrolysis, but this time by ATP molecules bound to the distant GroEL ring in the *trans*-position (Figure 6.15e,f). This is another example of the strong structural communication between the halves of the complex since ATP hydrolysis in the *trans*-GroEL ring affects the GroES binding site almost 100 Å away. Once GroES is released the polypeptide chain is released and it can bind to another GroEL–GroES complex to repeat the cycle. The native state is reached after iterative rounds involving multiple binding and release to GroEL–GroES complex.

The crucial question that remains to be answered is, of course, what happens to the polypeptide chain inside the closed *cis*-cavity. Two different

Petitioner Merck, Ex. 1014, p. 103

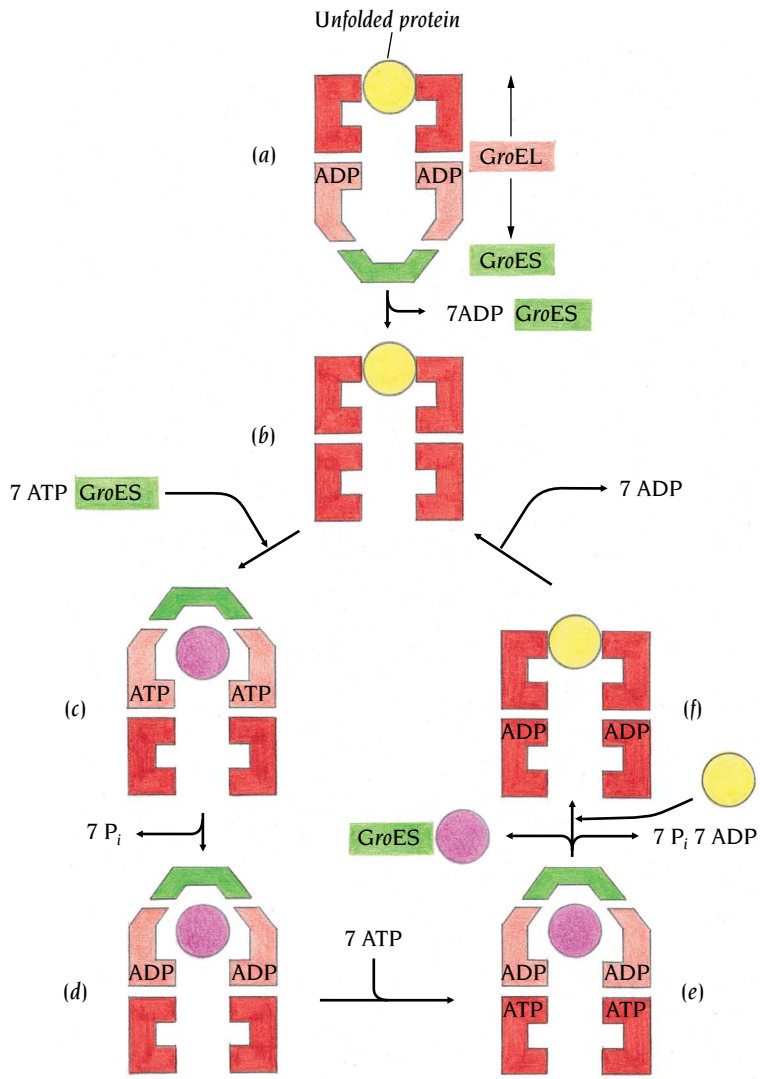


Figure 6.15 Possible functional cycle of the GroEL–GroES molecule. (a) An unfolded protein molecule (yellow) binds to one end of the GroEL–ADP complex (red) with bound GroES (green) at the other end. (b and c) GroES is released from the *trans*-position and rebound together with ATP at the *cis*-position (light red) of GroEL. (d) ATP hydrolysis occurs as the protein is folding or unfolding inside the central cavity. (e) ATP binding and hydrolysis in the *trans*-position is required for release of GroES and the protein molecule. (f) A new unfolded protein molecule can now bind to GroEL. (Adapted from M. Mayhew et al., *Nature* 379: 420–426, 1996.)

models have been proposed. Both envisage that unfolded or incorrectly folded proteins are recognized by their exposed hydrophobic areas and bound to hydrophobic regions inside the GroEL cavity. In one of these models the subsequent function of the cavity is to unfold unproductive intermediates and then eject them in the unfolded state into the bulk solution for spontaneous folding, giving them another chance to reach the folded state. In this model folding would occur in the solvent during jumps of the polypeptide between GroEL–GroES complexes. In the second model folding occurs inside the *cis*-cavity of the GroEL–GroES complex, either to the native state or to an intermediate state along one of the productive pathways to the native state. Experiments by the group of Ulrich Hartl, Memorial Sloan-Kettering Cancer Center, New York, have provided support, at least for some proteins, for the second model, which has the attractive feature that folding occurs in a closed environment preventing aggregation with other unfolded proteins during the folding process. However, GroEL-assisted folding of large protein molecules must be different since they are too large to fit inside the cavity. In addition, assisted folding by Hsp 70 does not occur inside a closed cavity.

The folded state has a flexible structure

For simplicity we have so far described a native folded protein molecule as being in one single state. However, within this state, the protein molecule does not have a static rigid structure at normal temperatures. Instead, all the atoms are subject to small temperature-dependent fluctuations. The molecule

Petitioner Merck, Ex. 1014, p. 104

as a whole undergoes **breathing** and every atom is constantly in motion. These atomic movements are in general random, but sometimes they can be collective and cause groups of atoms to move in the same direction. Side chains can flip from one conformation to another, some loop regions may not be fixed in one single conformation, helices may slide relative to each other and entire domains can change their packing contacts and open or close the distance between them. Usually these motions are small, a few tenths of an Ångstrom; but occasionally the collective motions can be large and very significant.

Such large collective movements are reflected in x-ray studies as a low level of electron density and in some cases no electron density at all. The regions that undergo these movements are usually described by crystallographers as flexible or disordered, but in order to distinguish between collective movements and a few discrete and well-ordered but different conformations of these regions, x-ray studies must be made at different temperatures. In NMR studies the experimental data for these regions are compatible with many different conformations. Insight into these individual and collective motions has been obtained by theoretical studies, called molecular dynamics simulations, which use classical Newtonian descriptions of atomic movements.

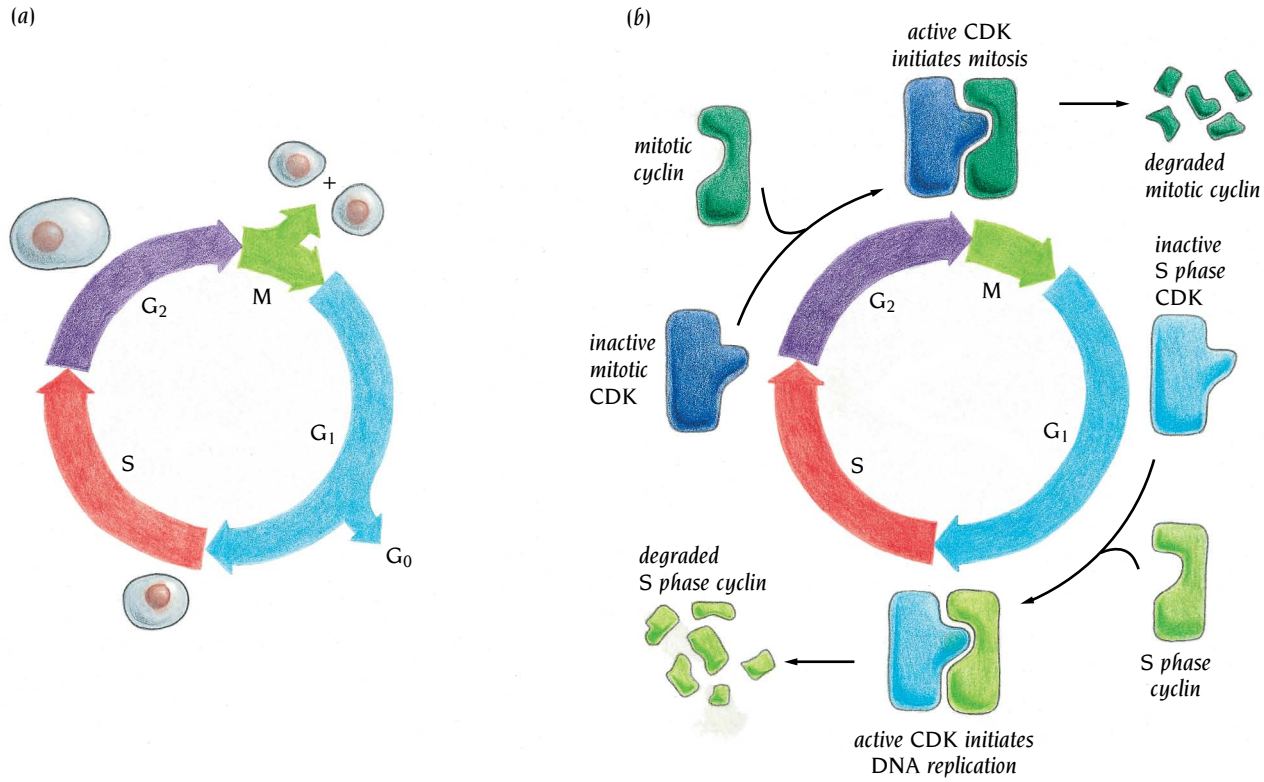
These calculations have shown that collective movements occur on the picosecond time scale for individual residues, and in nanoseconds for loop regions. Such movements are very important for the function of many protein molecules. Reactions such as electron transfer or ligand binding and release occur on these time scales and usually require movements of protein atoms. As soon as the structure of myoglobin was determined it was immediately apparent that the static picture of the myoglobin molecule seen in the crystal did not allow oxygen atoms to enter its binding site or diffuse out. We now know that while the myoglobin molecule is breathing, pathways are opened up between the solvent and the buried binding site to allow oxygen binding or release on a time scale of nanoseconds.

In addition to these small breathing movements of protein atoms there can also be larger conformational changes between different functional states of the molecule. Small differences in the environment such as different pH or the presence or absence of ligands can stabilize different conformational states of the protein. These conformational changes can vary from adjustments of side chain orientations in the active site to movements of loop regions, differences in relative orientation of domains or changes in the quaternary structure of oligomeric proteins. Such movements are usually essential for function, for enzyme catalysis, binding of antigens to antibodies, receptor–ligand interactions, muscle action and energy transduction and so on. We have already discussed two extreme examples, the effect of pH change on hemagglutinin (Chapter 5) and ATP hydrolysis in GroEL–GroES. We will here give a few further examples of striking conformational transitions.

Conformational changes in a protein kinase are important for cell cycle regulation

The cell division cycle of all eucaryotic cells can be divided into four major phases: G_1, S, G_2, and M (Figure 6.16a). In S phase (DNA synthesis), the entire DNA content of the nucleus is duplicated and the number of chromosomes doubled; in M phase (mitosis), the duplicated chromosomes of the parental cell are segregated using the microtubules of the mitotic spindle, such that each daughter cell receives the same set of chromosomes. After mitosis, including cytokinesis which divides the cytoplasm between the two daughter cells, and before the onset of the next S phase there is an interval designated Gap 1 or G_1. After the completion of DNA replication and chromosome duplication, in other words after S phase is completed, there is a second interval designated Gap 2, or G_2, before the onset of mitosis. The

Petitioner Merck, Ex. 1014, p. 105



complete cycle therefore is M, G_1, S, G_2 and M again. Throughout G_1, S and G_2 the cell's protein-synthesising machinery, macromolecules and organelles are built up, and the cell increases in volume. During mitosis the chromosomes and the cytoplasm are divided into two essentially equal parts. An additional stationary phase G_0 occurs in cells which are not actively dividing. How is the progression of a dividing cell through the consecutive phases of the cell cycle regulated? What triggers DNA synthesis in a G_1 cell and mitosis in a G_2 cell? Why don't the two daughter cells emerging from mitosis immediately begin a new round of DNA synthesis?

A combination of biochemical and genetic studies have shown that the progression through the cell cycle is dependent upon the successive activation of a series of enzymes called **cyclin-dependent protein kinases**, **CDKs** (Figure 6.16b). Each CDK during its transient existence phosphorylates target proteins that then directly or indirectly activate the next set of events of the cell cycle. Each CDK is a heterodimer comprising a catalytic subunit, the protein kinase, complexed with a **cyclin** molecule which activates the kinase. In vertebrate cells there are at least four different CDKs involved in control of passage through the cell cycle, as well as other CDKs with other functions. The different catalytic subunits are products of the genes of a closely related gene family. The different cyclins, one or more for each sort of catalytic subunit, are also members of a gene family. The cyclin components of the CDKs undergo sequential programmed synthesis, accumulation and then degradation; the short half-lives of the cyclins ensure that the CDKs of which they are part are active kinases only for short periods and at the correct time in the cell cycle. The CDKs can therefore act as a relay of switches, governing passage from G_1 to S phase, from G_2 to M phase and all other steps that constitute the cell cycle.

Although there is still a very great deal to learn about the physiological substrates of the CDKs and how the enzymes locate their targets, detailed structural information, including the activation of the kinase by cyclin, is available for CDK2-cyclin A, which regulates DNA replication in human cells. The x-ray structure of a functional fragment of cyclin A was determined by the group of Louise Johnson, Oxford University, that of inactive CDK2 by the group of Sung-Ho Kim, University of California, Berkeley, and that of the

Figure 6.16 (a) The five phases of a standard eucaryotic cell cycle. During M phase growth stops and the cell then divides. DNA replication is confined to the S phase; G_1 phase is the gap between M phase and S phase; G_2 phase is the gap between S phase and M phase. Cells which are not dividing enter the stationary phase, G_0. (b) The regulation of CDKs by cyclin degradation. Only two types of cyclin–CDK complexes are shown, one that triggers S phase and one that triggers M phase. In both cases the activation of CDK requires cyclin binding and its inactivation depends on cyclin degradation. (Adapted from B. Alberts et al, *Essential Cell Biology*, New York: Garland Publishing, 1998.)

Petitioner Merck, Ex. 1014, p. 106

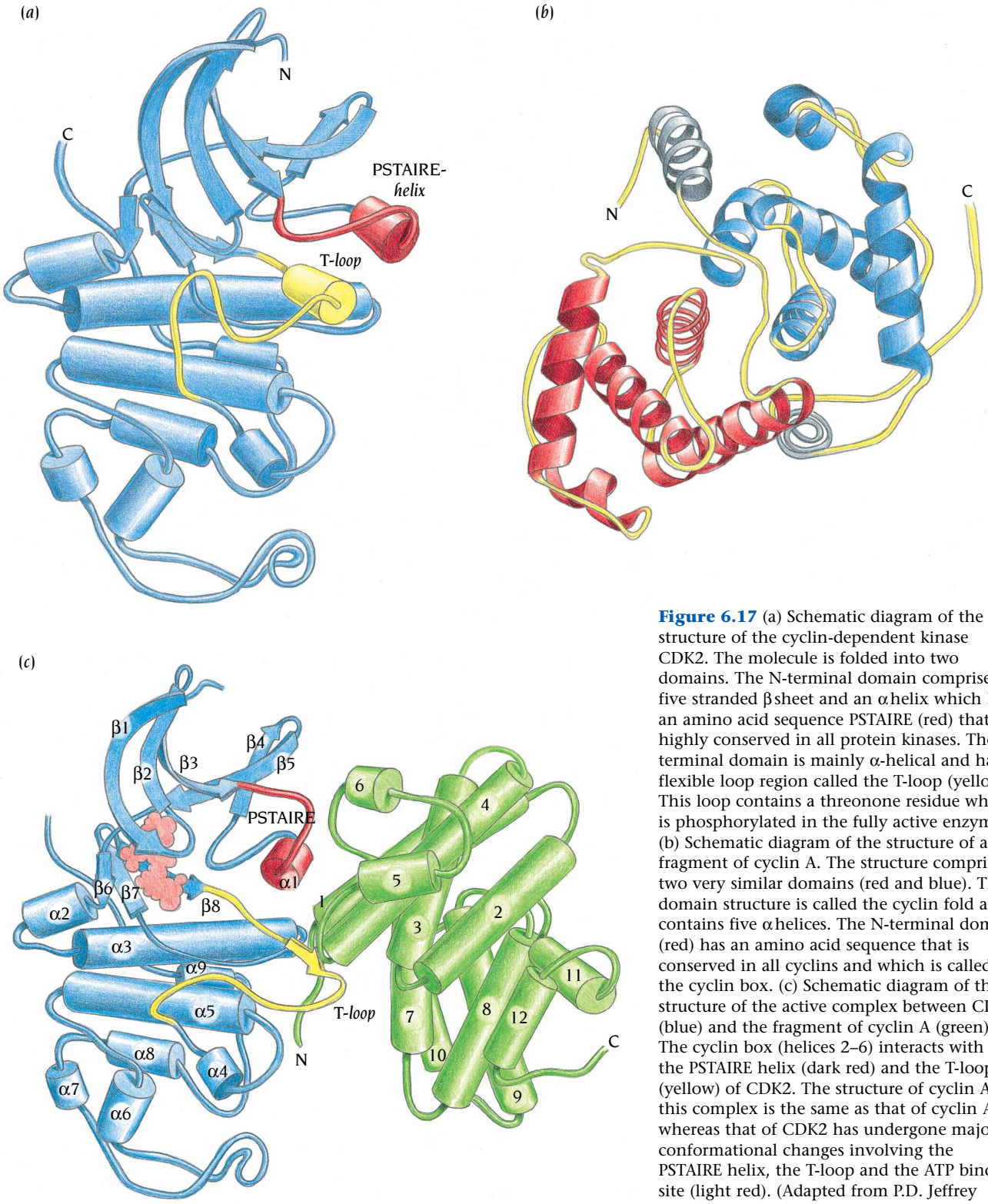


Figure 6.17 (a) Schematic diagram of the structure of the cyclin-dependent kinase CDK2. The molecule is folded into two domains. The N-terminal domain comprises a five stranded β sheet and an α helix which has an amino acid sequence PSTAIRE (red) that is highly conserved in all protein kinases. The C-terminal domain is mainly α-helical and has a flexible loop region called the T-loop (yellow). This loop contains a threonone residue which is phosphorylated in the fully active enzyme. (b) Schematic diagram of the structure of a fragment of cyclin A. The structure comprises two very similar domains (red and blue). This domain structure is called the cyclin fold and contains five α helices. The N-terminal domain (red) has an amino acid sequence that is conserved in all cyclins and which is called the cyclin box. (c) Schematic diagram of the structure of the active complex between CDK2 (blue) and the fragment of cyclin A (green). The cyclin box (helices 2–6) interacts with the PSTAIRE helix (dark red) and the T-loop (yellow) of CDK2. The structure of cyclin A in this complex is the same as that of cyclin A whereas that of CDK2 has undergone major conformational changes involving the PSTAIRE helix, the T-loop and the ATP binding site (light red). (Adapted from P.D. Jeffrey et al., *Nature* 376: 313–320, 1995.)

active cyclin A fragment-CDK2 complex by the group of Nicola Pavletich, Memorial Sloan-Kettering Cancer Center, New York. Comparison of these structures reveals how cyclin A binding to CDK2 causes large conformational changes in the active site of CDK2, converting the protein from an inactive to an active kinase. The structure of cyclin A, in contrast, is not changed.

CDK2 has two domains, a small (85 residue) amino-terminal domain comprising a single α helix and a five-stranded β sheet and a larger (213 residues) domain that is mainly α-helical (Figure 6.17a). The cofactor in the

Petitioner Merck, Ex. 1014, p. 107

phosphorylating reaction, ATP, is bound in a cleft between the two domains. The α helix in the small domain contains a sequence of residues, -Pro-Ser-Thr-Ala-Ile-Arg-Glu-, which is highly conserved in all protein kinases and which is called the PSTAIRE helix. Mutational studies have shown that the Glu residue in the PSTAIRE helix plays a crucial role for activity of the enzyme. The large, mainly α-helical domain has a flexible loop region, called the T-loop, which contains a threonine residue, Thr 160 in human CDK2, that is phosphorylated in the fully active enzyme. The cyclin A fragment (residues 173–432) that was used in these x-ray studies is built up from two domains with a very similar structure that is now called the cyclin fold (Figure 6.17b) and which comprises five α helices. The fragment activates CDK2 almost as well as the complete cyclin A molecule. The first domain has an amino acid sequence that is strongly conserved in all cyclins and which is called the cyclin box. In spite of the almost identical structures of the two domains their amino acid sequences are not similar, only one has the cyclin box.

In the cyclin A–CDK2 complex the cyclin box domain interacts with CDK2, mainly with the PSTAIRE helix and the T-loop (Figure 6.17c). The structure of cyclin A in the complex is virtually the same as that of cyclin A alone whereas that of CDK2 has undergone major conformational changes. The whole of the N-terminal domain has slightly changed its orientation relative to the C-terminal domain. In addition the PSTAIRE helix has moved closer to the active site cleft of CDK2 and rotated about 90° so that the catalytically essential residue Glu 51 points into the cleft instead of away from it in free CDK2 (Figure 6.18). Some of the main chain atoms of this helix have moved up to 8 Å due to these concerted movements. Coupled to the structural change of the PSTAIRE helix there is a major rearrangement of the T-loop with some residues moving up to 20 Å (Figure 6.19). As well as adopting a completely new position, the T loop undergoes a structural transformation. The part of the T-loop that is α-helical in free CDK2 melts away, and instead a β strand appears in the complex.

In free CDK2 the active site cleft is blocked by the T-loop and Thr 160 is buried (Figure 6.20a). Substrates cannot bind and Thr 160 cannot be phosphorylated; consequently free CDK2 is inactive. The conformational changes induced by cyclin A binding not only expose the active site cleft so that ATP and protein substrates can bind but also rearrange essential active site residues to make the enzyme catalytically competent (Figure 6.20b). In addition Thr

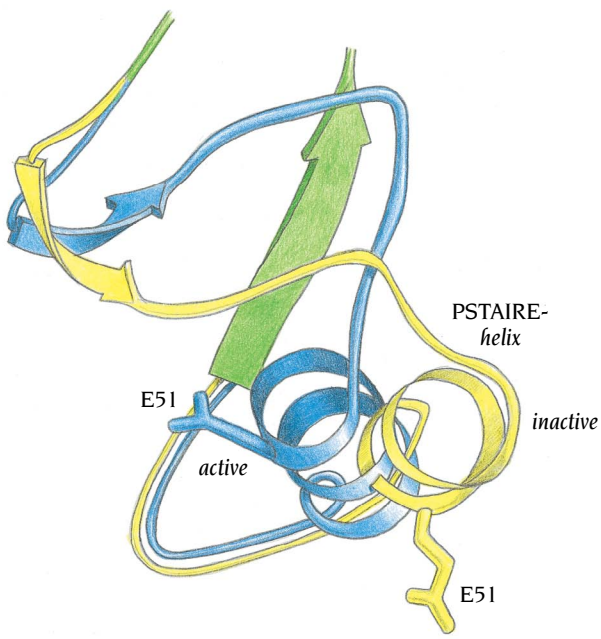


Figure 6.18 The PSTAIRE helix undergoes a major conformational change when CDK2 binds to cyclin A. In the inactive free CDK2 (yellow) the active site residue Glu 51 is far from the active site. Upon binding of cyclin A to CDK2 the PSTAIRE helix (blue) rotates 90° and changes its position so that Glu 51 becomes positioned into the active site. (Adapted from P.D. Jeffry et al., *Nature* 376: 313–320, 1995.)

Petitioner Merck, Ex. 1014, p. 108

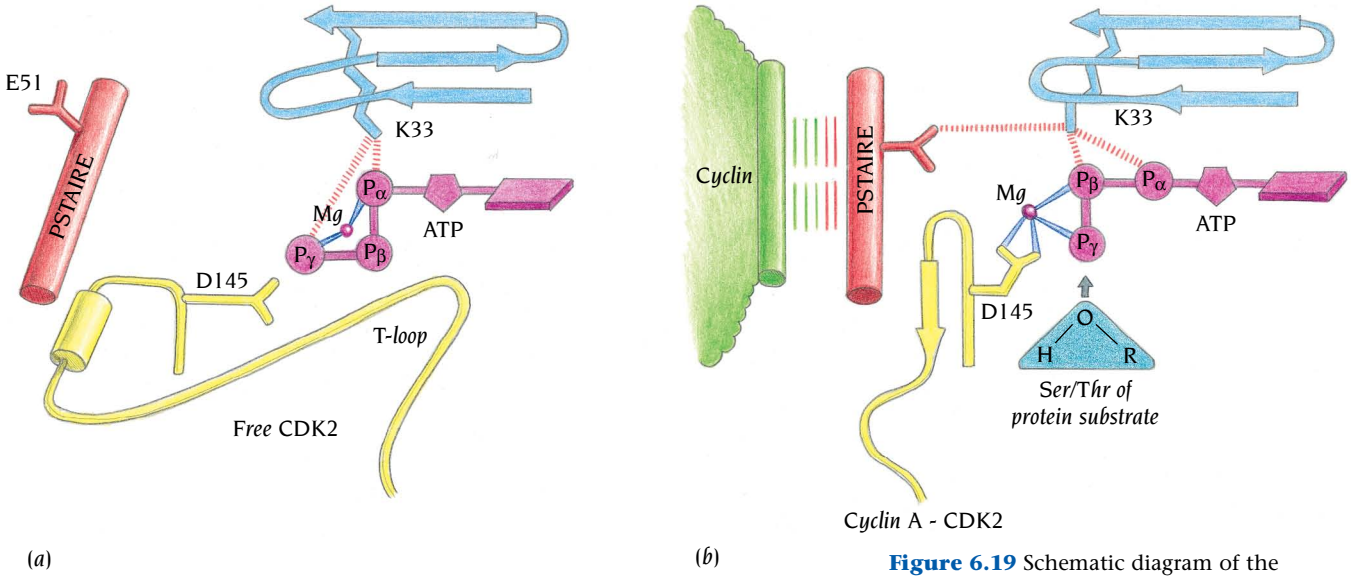


Figure 6.19 Schematic diagram of the conformational changes of CDK2 upon cyclin binding. (a) In the inactive form the PSTAIRE helix (red) is oriented such that Glu 51 points away from the ATP binding site (purple) and the T-loop (yellow) blocks the substrate binding site and prevents proteins from binding to CDK2. (b) In the active cyclin–CDK2 complex the PSTAIRE helix is reoriented so that Glu 51 points into the active site and forms a salt bridge to another residue involved in catalysis, Lys 33. The T-loop has drastically changed its conformation and one of its residues, Asp 145, forms ligands to the Mg atom in the active site. The substrate-binding site is now open, proteins can bind and the cyclin-CDK2 complex can phosphorylate Ser/Thr residues and thereby activate the bound proteins.

160 is exposed and ready to become phosphorylated, which enhances the catalytic activity. In short, spectacular flexibility of protein structure is essential for regulating the CDK family of enzymes and hence for controlling the cell cycle.

Peptide binding to calmodulin induces a large interdomain movement

Calmodulin is a ubiquitous calcium-binding protein of 148 amino acid residues that is involved in a range of calcium-dependent signaling pathways. Calmodulin binds to a variety of proteins such as kinases, calcium pumps and proteins involved in motility, thereby regulating their activities. The calmodulin-binding regions of these proteins, comprising about 20 sequentially adjacent residues, vary in their amino acid sequences but they all have a strong propensity to form α helices. Structure determinations of calmodulin alone and of complexes with peptides have shown that peptide binding induces a large conformational change in the calmodulin molecule.

The x-ray structure of free calmodulin was determined by the group of Charles Bugg, University of Alabama. It is a dumbbell-shaped molecule

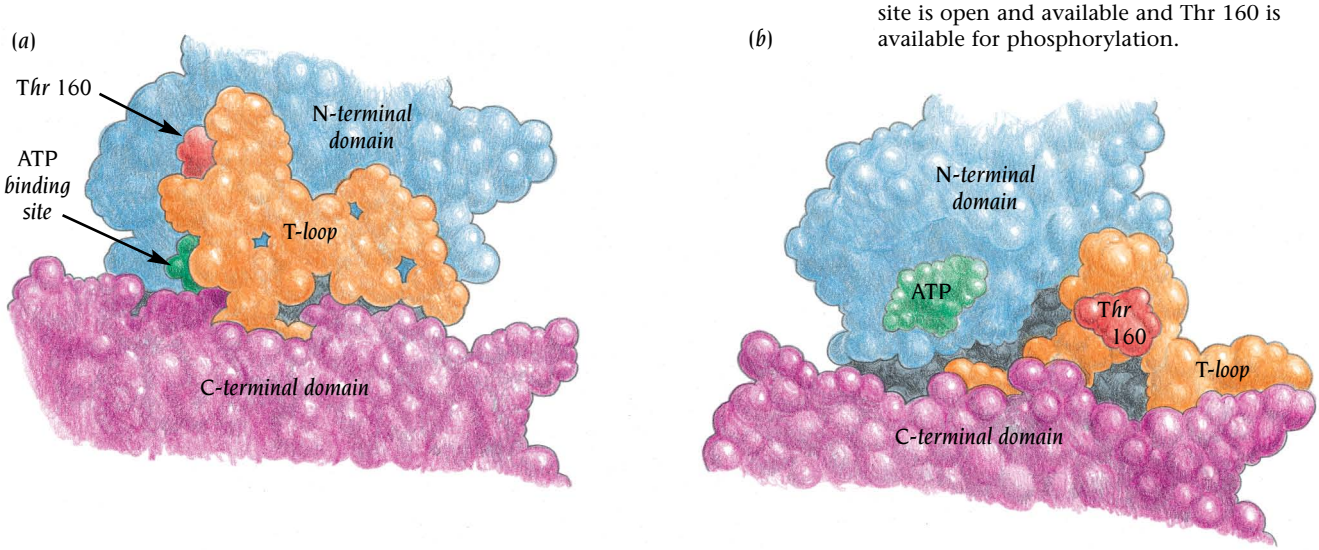


Figure 6.20 Space-filling diagram illustrating the structural changes of CDK2 upon cyclin binding. (a) The active site is in a cleft between the N-terminal domain (blue) and the C-terminal domain (purple). In the inactive form this site is blocked by the T-loop. (b) In the active cyclin bound form of CDK2 the T-loop has changed its structure, the active site is open and available and Thr 160 is available for phosphorylation.

Petitioner Merck, Ex. 1014, p. 109

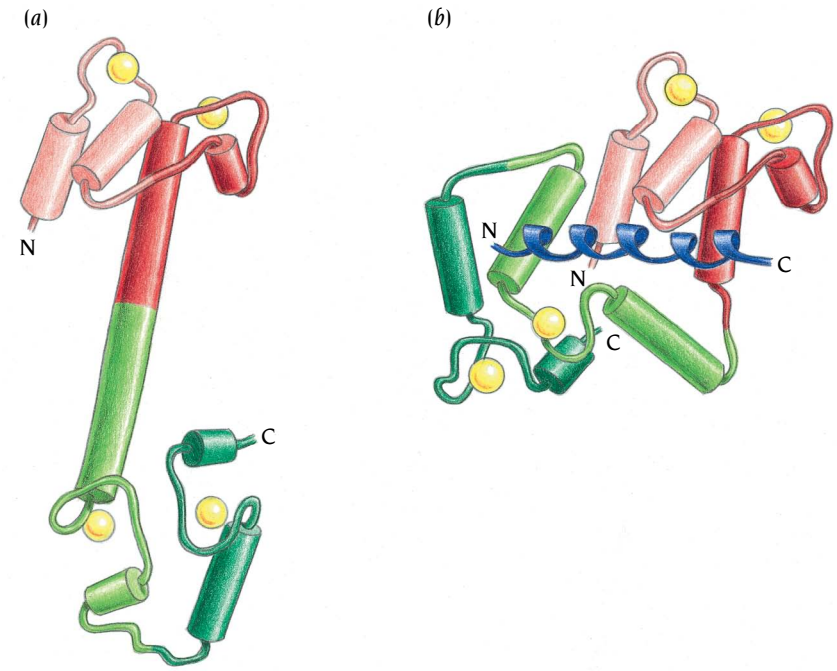


Figure 6.21 Schematic diagram of the conformational changes of calmodulin upon peptide binding. (a) In the free form the calmodulin molecule is dumbbell-shaped comprising two domains (red and green), each having two EF hands with bound calcium (yellow). (b) In the form with bound peptides (blue) the α helix linker has been broken, the two ends of the molecule are close together and they form a compact globular complex. The internal structure of each domain is essentially unchanged. The bound peptide binds as an α helix.

(Figure 6.21a) comprising two domains separated by a long straight α helix, similar in shape to troponin-C described in Chapter 2 (see Figure 2.13c). Each domain comprises two EF hands (see Figure 2.13a), each of which binds a calcium atom. The two domains are clearly separated in space at the two ends of the α helix linker.

The structures of two different complexes of calmodulin with binding peptides have been determined, one by the group of Ad Bax, National Institutes of Health, Bethesda, using NMR and the second by the group of Florante Quiocho, Baylor College of Medicine, Houston, using x-ray crystallography. These two structures are quite similar but the molecular shape of calmodulin in these complexes is very different from that of free calmodulin. The internal structures of the two domains have not changed but the α helix linker has been broken into two helices that are oriented in different directions so that the relative positions of the domains have changed (Figure 6.21b). They are now close together forming a compact molecule of ellipsoidal shape. The bound peptide is in a wide cleft between the two domains and adopts an α-helical conformation.

When calmodulin binds a ligand, only five groups actually change their conformation. They are five consecutive residues in the linker helix, which unwind and turn into a loop region. The α helix continues after this loop but now in an entirely new direction, which positions the second domain close to the first and in a different orientation. This rather small local change in peptide conformation causes one of the largest ligand-induced interdomain motions known in a protein, comparable to the large repositioning of domains during the pH-induced conformational change of hemagglutinin discussed in Chapter 5.

Serpins inhibit serine proteinases with a spring-loaded safety catch mechanism

Infections in the lung elicit an accumulation of activated leucocytes that secrete enzymes involved in removing the damage done by the infection. The most important of these enzymes is neutrofil elastase, which belongs to the serine proteinase family of enzymes described in Chapter 11. The health of the lung depends to a large extent on proper control of the activity of this enzyme, which is achieved by a blood plasma proteinase inhibitor named, misleadingly, **α_1-antitrypsin** because it also inhibits other serine proteinases,

Petitioner Merck, Ex. 1014, p. 110

among them trypsin. Alpha$_1$-antitrypsin belongs to a family of serine proteinase inhibitors found in blood plasma that are collectively called **serpins**. Other members of this family are **antithrombin** and **plasminogen activator inhibitor**, **PAI**, both of which are essential regulators of the blood coagulation cascade of reactions. All serpin molecules are homologous with very similar three-dimensional structures.

Serpins form very tight complexes with their corresponding serine proteinases, thereby inhibiting the latter. A flexible loop region of the serpin binds to the active site of the proteinases. Upon release of the serpin from the complex its polypeptide chain is cleaved by the proteinase in the middle of this loop region and the molecule is subsequently degraded. In addition to the active and cleaved states of the serpins there is also a latent state with an intact polypeptide chain that is functionally inactive and does not bind to the proteinase.

The structures of all three states of the serpins have been determined by x-ray crystallography, the cleaved form of α_1-antitrypsin by the group of Robert Huber, Max-Planck Institute for Biochemistry, Munich, the latent form of PAI by the group of Elizabeth Goldsmith, University of Texas, Dallas, and the active form of antithrombin by the groups of Wim Hol, University of Washington, Seattle, and of Robin Carrel, Cambridge University. In addition, the group of Robin Carrel has determined the structure of another member of the serpin family, **ovalbumin**, which is present in uncleaved form in egg white. The general folds of these serpin molecules in their different states are the same, but the positions of the flexible loop regions vary in a novel and intriguing way.

The serpin fold comprises a compact body of three antiparallel β sheets, A, B and C, which are partly covered by α helices (Figure 6.22). In the structure of the uncleaved form of ovalbumin, which can be regarded as the canonical structure of the serpins, sheet A has five strands. The flexible loop starts at the end of strand number 5 of β sheet A (β15 in Figure 6.22), then

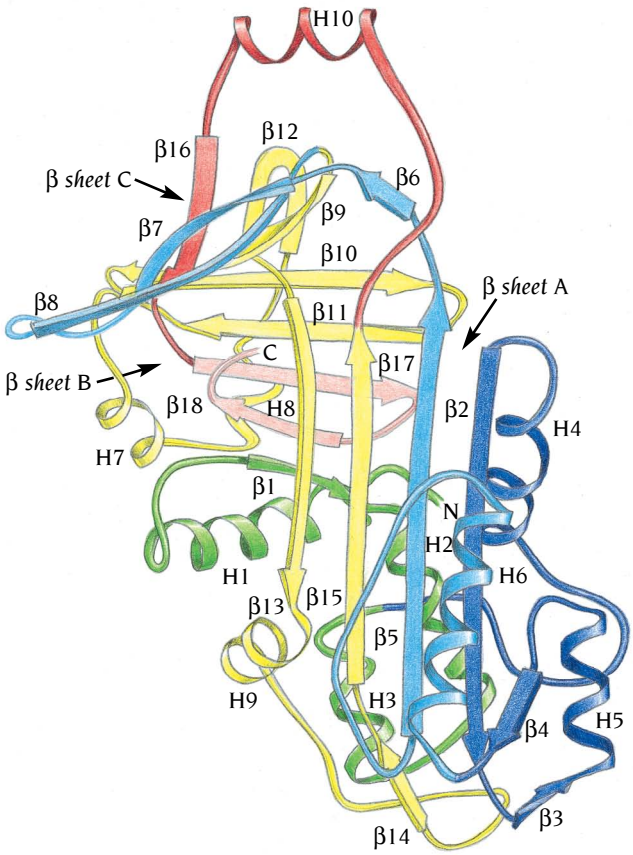


Figure 6.22 Schematic diagram of the structure of ovalbumin which illustrates the serpin fold. The structure is built up of a compact body of three antiparallel β sheets, A, B, and C, surrounded by α helices. The polypeptide chain is colored in sections from the N-terminus to facilitate following the chain tracing in the order green, blue, yellow, red and pink. The red region corresponds to the active site loop in the serpins which in ovalbumin is protruding like a handle out of the main body of the structure. (Adapted from R.W. Carrell et al., *Structure* 2: 257–270, 1994.)

Petitioner Merck, Ex. 1014, p. 111

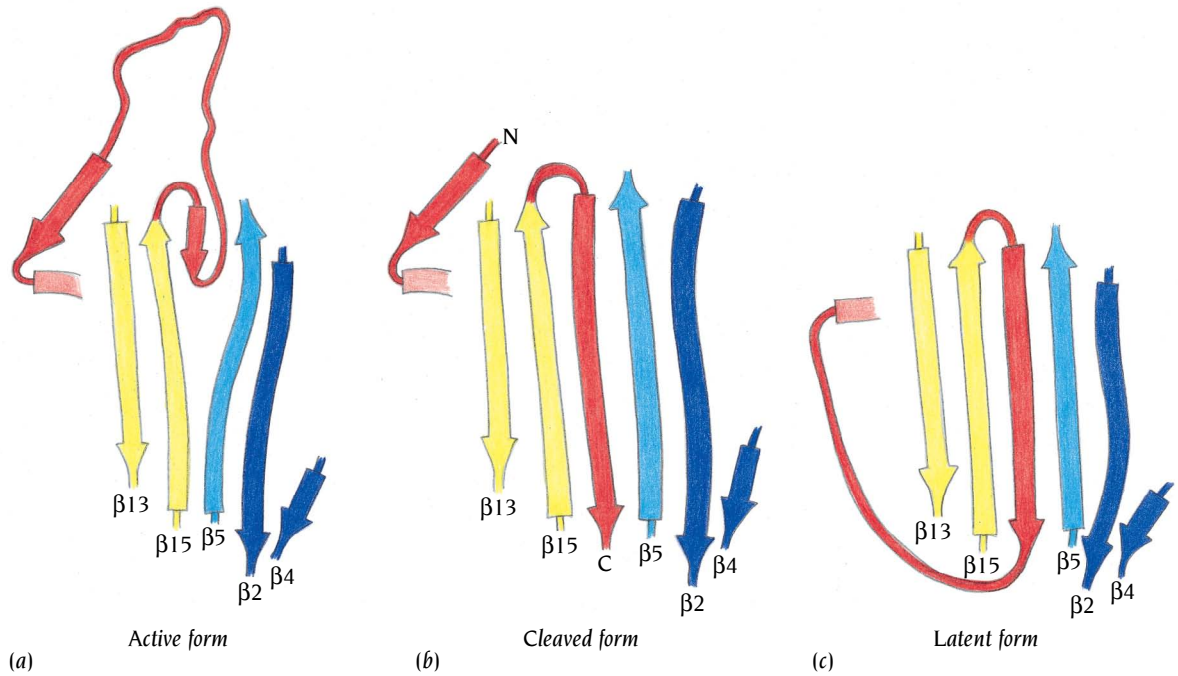


Figure 6.23 Schematic diagram illustrating the active site loop regions (red) in three forms of the serpins. (a) In the active form the loop protrudes from the main part of the molecule poised to interact with the active site of a serine proteinase. The first few residues of the loop form a short β strand inserted between β5 and β15 of sheet A. (b) As a result of inhibiting proteases, the serpin molecules are cleaved at the tip of the active site loop region. In the cleaved form the N-terminal part of the loop inserts itself between β strands 5 and 15 and forms a long β strand (red) in the middle of the β sheet. (c) In the most stable form, the latent form, which is inactive, the N-terminal part of the loop forms an inserted β strand as in the cleaved form and the remaining residues form a loop at the other end of the β sheet. (Adapted from R.W. Carrell et al., *Structure* 2: 257–270, 1994.)

forms an α helix outside the top of the molecule followed by an edge strand of β sheet C (β16) and finally ends at the beginning of one of the strands in β sheet B (β17). The central α-helical region of the loop, which contains the cleavage site for the serpins, is extended like a handle on the outside of the molecule.

The flexible loop region in the active form of antithrombin (Figure 6.23a) is in the same general position as in ovalbumin but the first few residues form a short sixth β strand in β sheet A inserted between strands β5 and β15. Furthermore there is no α helix in the loop which is extended outside the main body of the molecule, ready to be inserted into the active site of thrombin.

In the cleaved form of α_1-antitrypsin the first half of the loop region up to the cleavage site forms a complete β strand inserted between strands β5 and β15 in β sheet A (Figure 6.23b). The other half of the loop region has approximately the same position as in the active form of antithrombin. The two new ends of the polypeptide, which are joined in the active form, are here at opposite ends of the molecule, 70 Å apart. Finally, in the latent form of PAI the additional β strand in β sheet A is present as in the cleaved form of α_1-antitrypsin, but the rest of the flexible loop region makes a loop on the outside of the molecule and enters the β strand of β sheet B without forming an edge strand of β sheet C (Figure 6.23c).

The conversion of the active form to the latent form involves the conversion of a loop into a long β strand inserted in the middle of a β sheet. To achieve the remarkable structural change of inserting a β strand in the middle of a stable preformed β sheet, adjacent strands in the β sheet must first be separated. This involves breaking many hydrogen bonds as well as changing a number of hydrophobic packing contacts in the interior of the molecule. New hydrogen bonds and new packing contacts must then be made when the extra β strand is inserted. Such major changes in a β structure were quite unforseen before these serpin structures were determined and have not yet been observed in any other system.

Which of these forms is most stable? Surprisingly, the active form is less stable than the latent form. Conversion from the active to the latent form can occur spontaneously over a period of hours or days *in vitro* and more quickly under mild denaturing conditions. In contrast, recovery of the active form from the latent form requires complete unfolding of the latent form

Petitioner Merck, Ex. 1014, p. 112

under strong denaturing conditions, and subsequent refolding. Refolding does, however, produce the less stable active form in preference to the more stable latent form. This is one of the few pieces of direct experimental evidence that the folding process can be kinetically controlled through intermediates that produce a native state that is not the thermodynamically most stable state.

In vivo PAI and antithrombin are stabilized in their active forms by binding to vitronectin and heparin, respectively. These two serpins seem to have evolved what Max Perutz has called "a spring-loaded safety catch" mechanism that makes them revert to their latent, stable, inactive form unless the catch is kept in a loaded position by another molecule. Only when the safety catch is in the loaded position is the flexible loop of these serpins exposed and ready for action; otherwise it snaps back and is buried inside the protein. This remarkable biological control mechanism is achieved by the flexibility that is inherent in protein structures.

Emphysema is often associated with a specific mutation of the serpin antitrypsin. The mutant serpin molecules form aggregates in the liver, causing a deficiency of antitrypsin in the blood plasma and consequently increased proteolytic degradation of elastin fibers in the lung by the enzyme elastase. It has been shown that the formation of aggregates *in vivo* is due to an extremely slow folding process of the mutant antitrypsin leading to accumulation of a folding intermediate that aggregates. This is one example of aggregation of incompletely folded or misfolded molecules that can lead to pathologic consequences or even severe disease. Other examples involve the formation of large aggregates, plaques, of proteins in amyloid structures associated with Alzheimer's disease and spongiform encephalopathies such as scrapie, BSE "mad cow disease" and Creutzfeld-Jacob disease (see Chapter 14). A better understanding of the folding and misfolding processes might, therefore, open up new approaches to drugs for these diseases.

Effector molecules switch allosteric proteins between R and T states

In 1963 Jaques Monod, Jean-Pierre Changeaux and Francois Jacob published a theory that radically changed our views on the control of protein function and which still influences protein biochemists as well as structural biologists. Their theory of **allosteric control** provided a unifying theme for such diverse concepts as **feedback inhibition** of enzymes, repressor/corepressor binding (see Chapter 8) and **cooperative binding** of ligand by proteins, with oxygen binding to hemoglobin as the prime example. The allosteric theory they proposed has the following main features. Cooperative substrate binding and modification of a protein's activity by allosteric effector molecules may arise in proteins with two or more preformed structural states that are in equilibrium. Substrates and effector molecules bind at different sites on the protein and therefore need no stereochemical relationship to each other, hence the name allostery (different shapes).

The theory predicts that such proteins are built up of several subunits which are symmetrically arranged and that the two states differ by the arrangements of the subunits and the number of bonds between them. In one state the subunits are constrained by strong bonds that would resist the structural changes needed for substrate binding, and this state would consequently bind substrates weakly; they called it the tense or T state. In the other state, called the R state, these constraints are relaxed.

This **concerted model** assumes furthermore that the symmetry of the molecule is conserved so that the activity of all its subunits is either equally low or equally high, that is, all structural changes are concerted. Subsequently Daniel Koshland, University of California, Berkeley, postulated a **sequential model** in which each subunit is allowed independently to change its tertiary structure on substrate binding. In this model tertiary structural changes in the subunit with bound ligand alter the interactions of this

Petitioner Merck, Ex. 1014, p. 113

subunit with its neighbours and this leads sequentially to changes in the latter's reactive sites. Koshland's model was based on his theories of induced fit; ligand binding induces a conformational change that converts an enzyme from an inactive to an active state.

For many years hemoglobin was the only allosteric protein whose stereochemical mechanism was understood in detail. However, more recently detailed structural information has been obtained for both the R and the T states of several enzymes as well as one genetic repressor system, the trp-repressor, described in Chapter 8. We will here examine the structural differences between the R and the T states of a key enzyme in the glycolytic pathway, phosphofructokinase.

X-ray structures explain the allosteric properties of phosphofructokinase

Phosphofructokinase, **PFK**, is the key regulatory enzyme in glycolysis, the breakdown of glucose to generate ATP which occurs in most cells (Figure 6.24a). The enzyme catalyzes one of the early steps in this pathway, phosphorylation by ATP of fructose-6-phosphate, F6P, to fructose-1,6-bisphosphate (Figure 6.24b). Binding to PFK of one of the substrate molecules, F6P, is highly cooperative whereas binding of the second substrate ATP is

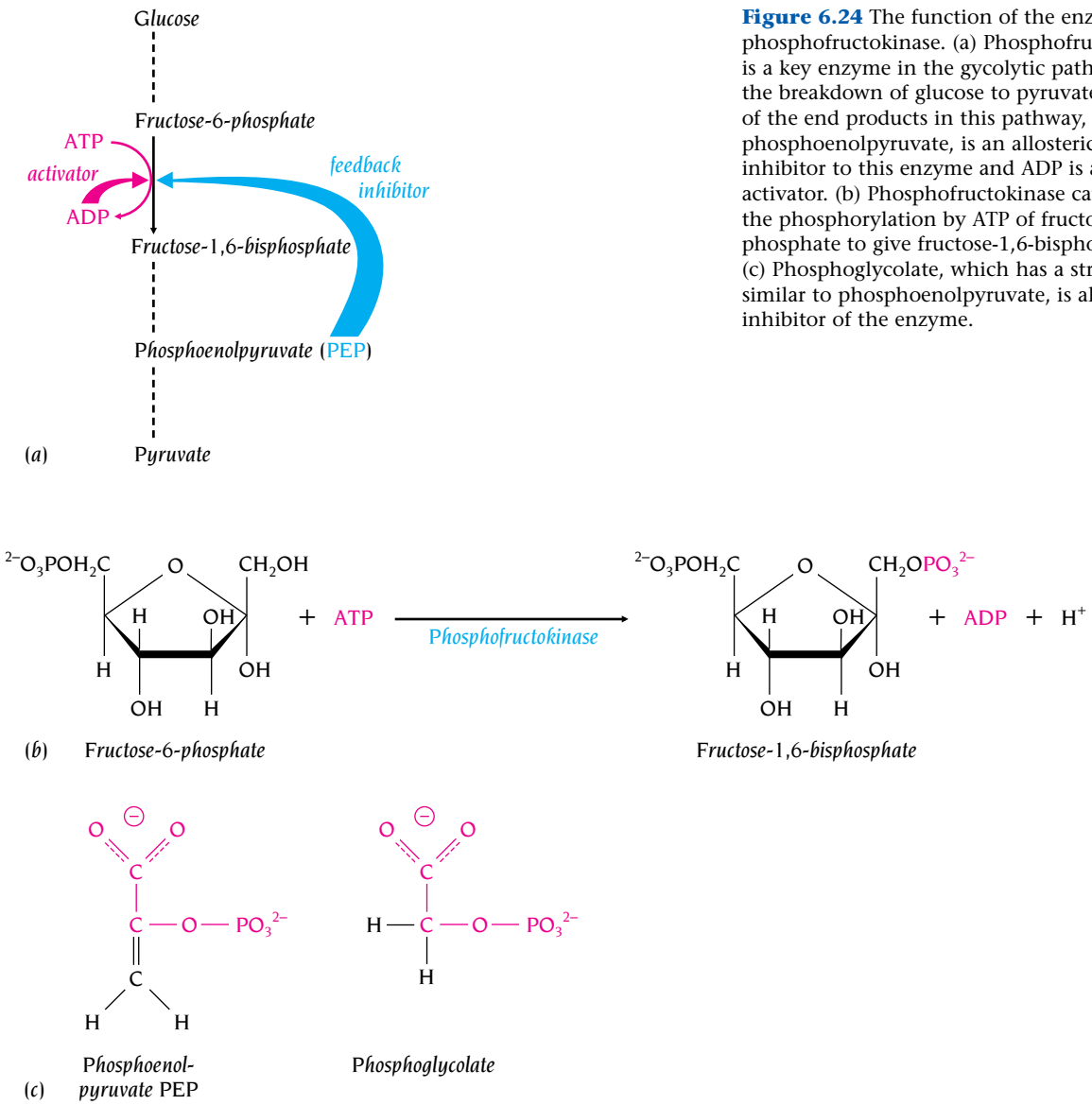


Figure 6.24 The function of the enzyme phosphofructokinase. (a) Phosphofructokinase is a key enzyme in the gycolytic pathway, the breakdown of glucose to pyruvate. One of the end products in this pathway, phosphoenolpyruvate, is an allosteric feedback inhibitor to this enzyme and ADP is an activator. (b) Phosphofructokinase catalyzes the phosphorylation by ATP of fructose-6-phosphate to give fructose-1,6-bisphosphate. (c) Phosphoglycolate, which has a structure similar to phosphoenolpyruvate, is also an inhibitor of the enzyme.

Petitioner Merck, Ex. 1014, p. 114

noncooperative. Phosphofructokinase is inhibited by phosphoenolpyruvate, PEP, one of the products of a late step in the glycolytic pathway, and by chemical analogs of PEP, for example 2-phosphoglycolate (Figure 6.24c). By contrast the reaction rate is enhanced by ADP, an allosteric effector molecule. The regulation of PFK by effector molecules is the main way that the glucose degradation by glycolysis is controlled in cells.

Because of the crucial role of this enzyme in one of the most important biochemical pathways in the cell, its allosteric properties have been studied extensively in solution. Interpretation of these studies in terms of the theory of allosteric enzymes led Monod and coworkers to conclude that:

1. The enzyme is made of four identical subunits each having a single binding site for each ligand.
2. The subunits can switch between two distinct conformational states, R and T, which are in equilibrium.
3. The transitions between these states in each tetrameric molecule are concerted, in other words all four subunits of each molecule are in the same state, either R or T.
4. The two states have the same affinity for ATP but differ with respect to their affinity for the substrate F6P, the allosteric effector ADP and the inhibitor PEP. Because of these differences in affinity, ligand binding can shift the equilibrium between the R and T states to favor one or the other state depending on which ligand is bound.

The group of Phil Evans, MRC Laboratory of Molecular Biology, Cambridge, UK, has determined x-ray structures of bacterial PFK both in the R and the T states. These studies have confirmed the above conclusions and given insight into how an allosteric enzyme accomplishes its complex behavior.

Each subunit of the homotetrameric PFK of *Escherichia coli* comprises 320 amino acids arranged in two domains, one large and one smaller, both of which have an α/β structure reminiscent of the Rossman fold (Figure 6.25).

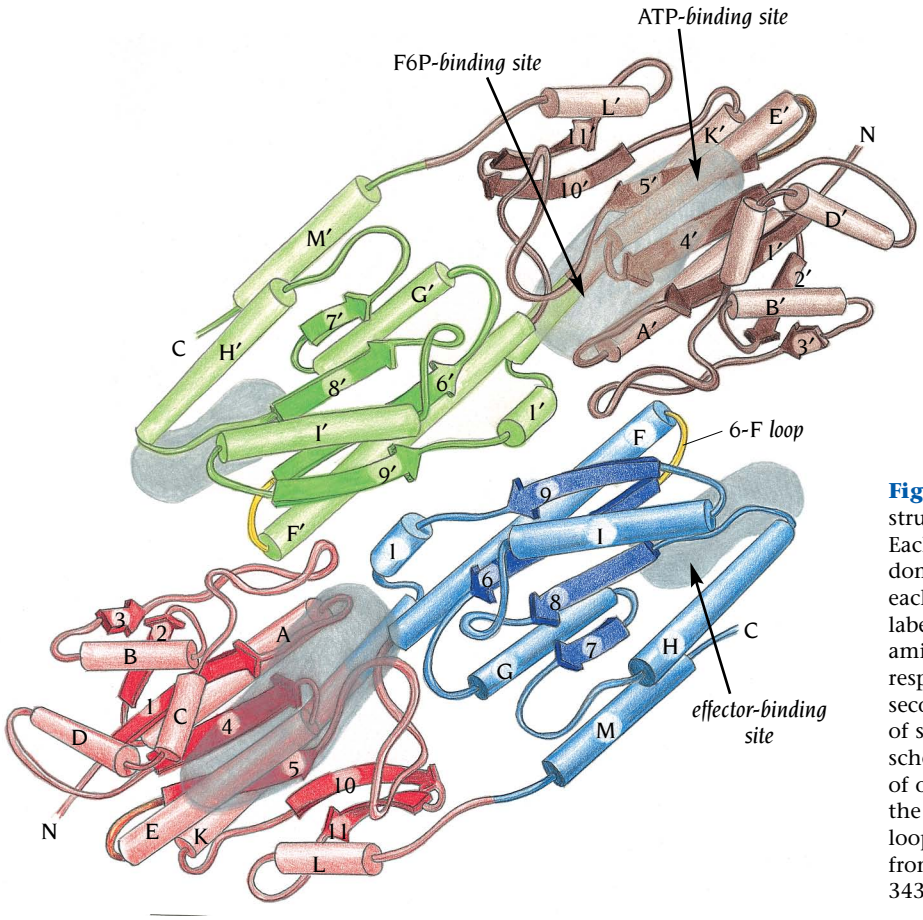


Figure 6.25 Schematic diagram of the structure of one dimer of phosphofructokinase. Each polypeptide chain is folded into two domains (blue and red, and green and brown), each of which has an α/β structure. Helices are labeled A to M and β strands 1 to 11 from the amino terminus of one polypeptide chain, and respectively A′ to M′ and 1′ to 11′ for the second polypeptide chain. The binding sites of substrate and effector molecules are schematically marked in gray. The effector site of one subunit is linked to the active site of the other subunit of the dimer through the 6-F loop between helix F and strand 6. (Adapted from T. Schirmer and P.R. Evans, *Nature* 343: 140–145, 1990.)

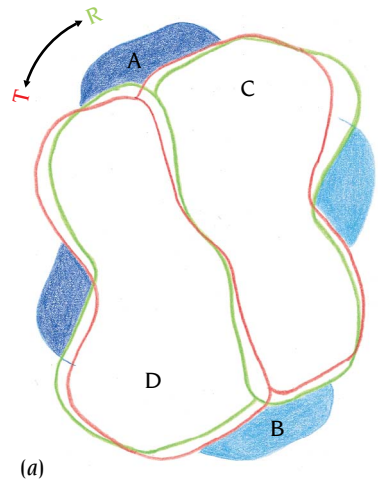
Petitioner Merck, Ex. 1014, p. 115

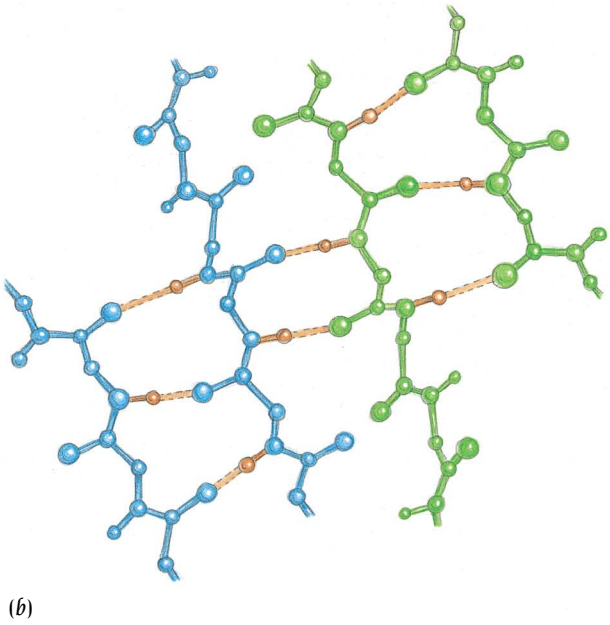
The subunits are pairwise linked into two dimers (A-B and C-D in Figure 6.26a) with extensive close contacts between the subunits within the dimers. The two dimers are loosely packed against each other into a symmetrical tetramer. The close contacts between the two subunits of the dimer are the same in the R and the T states but the interactions between the dimers are quite different. The orientation of the dimers with respect to each other in the R and T states differs by a rotation of 7°. This difference affects the packing of the dimers against each other and hence the quaternary structure. In the T state the dimers are close together and there are direct hydrogen bonds between two β strands, one from each dimer (Figure 6.26b). In the R state these two β strands are further apart and the gap between them is filled with water molecules that form hydrogen bond bridges (Figure 6.26c). The inclusion or exclusion of water molecules between the dimers is an all-or-none effect that acts like a two-way switch. As we shall see this change in quaternary structure of the tetrameric molecule is intimately linked to differences in tertiary structure of the subunits in the R and T states.

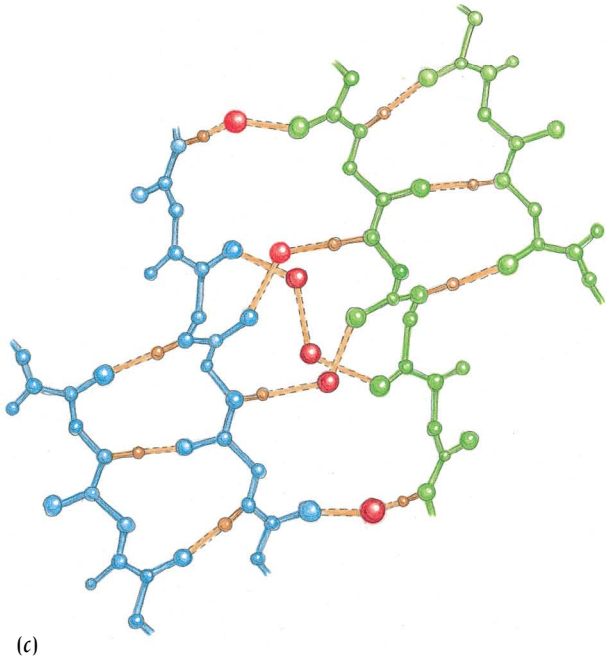
Each of the four subunits contains three binding sites (see Figure 6.25). There is one site for each of the substrate molecules, ATP and F6P, which together form the active site facing a cleft between the two domains. The third binding site of each subunit is the regulator binding site to which both the inhibitor PEP and the allosteric activator ADP can bind; this site is distant from the active site. Evans has studied crystals of the catalytically competent R state of three types: (1) with the subunits complexed with substrates, (2) subunits complexed with ADP, and (3) subunits unliganded. He also studied crystals of the T state in which the subunits are complexed with the inhibitor 2-phosphoglycolate.

The catalytic site of each subunit is in the cleft between the two domains. The large domain binds ATP with the terminal phosphate pointing into the cleft. The main binding site for the second substrate molecule, F6P, is in the smaller domain and the phosphate group of F6P interacts with a neighboring subunit affecting subunit interactions that are crucial for catalytic activity. In the active R state this phosphate group forms hydrogen bonds to an arginine residue, Arg 162, of a small α helix, called the 6-F helix, in the neighboring subunit (Figure 6.27a). By contrast in the T form this helix is unwound and instead forms a loop with Arg 162 pointing away from the F6P molecule (Figure 6.27b). In the T state a negatively charged glutamate side chain, Glu 161, occupies the same position as the positively charged Arg 162 in the R state. This negatively charged glutamate 161 repels the negative charge of the phosphate group of F6P. Consequently, when the R state is transformed to

Figure 6.26 The quaternary structure of phosphofructokinase. (a) The four subunits are pairwise arranged in two dimers A-B (blue) and C-D (red or green). The subunit interactions within the dimers are extensive and tight whereas the two dimers are loosely packed against each other and the packing contacts are different in the R and the T states. The orientation of the dimers with respect to each other in the T (red contours of the C-D dimer) and R (green contours) differs by a rotation of 7°. (b) The dimers are close together in the T state and there are direct hydrogen bonds between two β strands, one from the A-B dimer (blue) and one from the C-D dimer (green). Hydrogen bonds are shown in orange. (c) The dimers are further apart in the R state and there is a gap between the two β strands from the two dimers which is filled by water molecules (red). These water molecules form bridges between the dimers by making hydrogen bonds to the C=O and N-H groups of the two β strands.







Petitioner Merck, Ex. 1014, p. 116

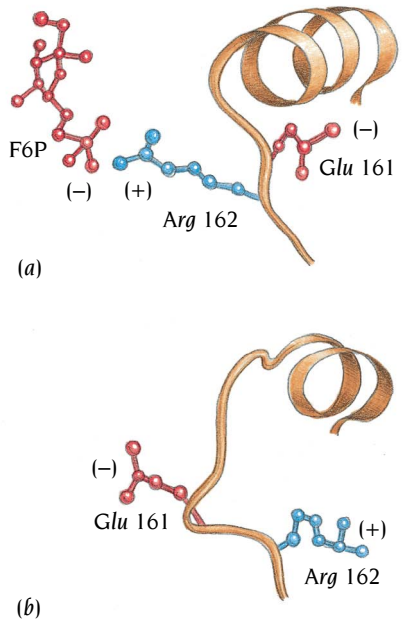


Figure 6.27 Conformational changes in the active site of phosphofructokinase. (a) In the active R state the phosphate group of the substrate fructose-1-phosphate, F6P, (red) forms a salt bridge to an arginine residue, Arg 162, of a small α helix (orange). This salt bridge contributes substantially to promote binding of the substrate to the enzyme. (b) In the inactive T state the helix has been partially unwound and changed its orientation so that Arg 162 points away from the substrate binding site. Instead a negatively charged glutamate residue, Glu 161, points towards the phosphate binding site of the substrate molecule. Repulsive forces between the negative charges of Glu 161 and the phosphate of F6P prevent binding and result in a thousandfold lower affinity for F6P when the enzyme is in the T state compared with the R state.

the T state the change in the F6P binding site results in a thousandfold lower affinity for F6P. In addition, catalytically important residues are properly arranged for catalysis in the R state but not in the T state. Since the binding site of ATP is virtually identical in the R and T states the binding of ATP is not affected by the structural differences between them.

The change in the quaternary structure and the structural change in the 6-F helix as the molecule moves from one state to the other are intimately related. The dimer interactions in the T state are not compatible with the presence of the 6-F helix, which would, if present, clash with the neighbouring dimer. The quaternary structure of the T state requires that the 6-F helix be unwound. Conversely the R state quaternary structure depends on the presence of the 6-F helix.

The basic kinetic properties of this allosteric enzyme are clearly explained by combining Monod's theory and these structural results. The tetrameric enzyme exists in equilibrium between a catalytically active R state and an inactive T state. There is a difference in the tertiary structure of the subunits in these two states, which is closely linked to a difference in the quaternary structure of the molecule. The substrate F6P binds preferentially to the R state, thereby shifting the equilibrium to that state. Since the mechanism is concerted, binding of one F6P to the first subunit provides an additional three subunits in the R state, hence the cooperativity of F6P binding and catalysis. ATP binds to both states, so there is no shift in the equilibrium and hence there is no cooperativity of ATP binding. The inhibitor PEP preferentially binds to the effector binding site of molecules in the T state and as a result the equilibrium is shifted to the inactive state. By contrast the activator ADP preferentially binds to the effector site of molecules in the R state and as a result shifts the equilibrium to the R state with its four available, catalytically competent, active sites per molecule.

Conclusion

The thermodynamic stability of a protein in its native state is small and depends on the differences in entropy and enthalpy between the native state and the unfolded state. From the biological point of view it is important that this free energy difference is small because cells must be able to degrade proteins as well as synthesize them, and the functions of many proteins require structural flexibility.

When a fully extended unfolded polypeptide chain begins to fold, hydrophobic residues tend to be buried in the interior, greatly restricting the number of possible conformations the chain can assume, and therefore allowing proteins to fold in seconds rather than years. Within milliseconds the polypeptide chain achieves the molten globule state, a term used to describe a set of structures that have in common a loosely packed hydrophobic core and some secondary structure. Some proteins have one preferred folding pathway, while others seem to have multiple parallel pathways to the native state. There are certain high energy barriers to folding such as, for

Petitioner Merck, Ex. 1014, p. 117

example, the formation of correct disulfide bonds and the isomerization of proline residues. Cells contain enzymes such as protein disulfide isomerases and *cis–trans*–proline isomerases that catalyze these reactions and therefore overcome what otherwise would be an insuperable energy barrier to rapid folding.

The cytoplasm of all cells contains folded proteins and folding polypeptides at high concentrations. Unfolded proteins with exposed hydrophobic patches aggregate easily by non-specific hydrophobic interactions. To circumvent this problem a class of proteins called chaperones have evolved to sequester unfolded polypeptides. The complex structure of one class of multimeric chaperones, the chaperonins GroEL and GroES, has been elucidated and has shed light on how chaperones function. These chaperonins are short cylinders which, because they have hydrophobic residues in the interior, can bind to any unfolded polypeptide that has large exposed hydrophobic patches, regardless of its amino acid sequence. Once the polypeptide chain is shielded inside the chaperonins it is protected from aggregation with other protein molecules. During folding some polypeptide chains go through many cycles of binding and release from the chaperonins.

Even inside crystals all atoms in protein molecules undergo small oscillations. Protein structures determined by x-rays are an average of these breathing structures. In addition to breathing, some proteins undergo large conformational changes in response to ligand binding or to changes in their environment, and these conformational changes are essential for function. The switches that control the successful passage through the eucaryotic cell cycle depend on changes in the conformation of an α helix and a flexible loop region in the cyclin-dependent kinases. In the case of calmodulin, structural changes play a crucial role in calcium signaling pathways. The dumbbell-shaped inactive molecule collapses into a globular structure that binds to regulatory proteins in such pathways. The serpins, a class of specific serine proteinase inhibitors, undergo an extraordinary conformational change: activation of the inactive form involves conversion of a β strand in the middle of a β sheet into a flexible loop and the converse occurs when the active form changes into the latent form.

Many multimeric enzymes and some other multimeric proteins, the classic examples being hemoglobin and phosphofructokinase, PFK, are subject to allosteric control. Allosteric proteins exist in two states, classically known as the R (relaxed) and the T (tense) states. Effector molecules have a high affinity for only one of these states. Therefore when an effector molecule is present it shifts the equilibrium to favor the high affinity state. The binding site for effector molecules is unrelated to and distinct from the active site. In the case of PFK there are two effector molecules, the activator ADP, which shifts the four identical subunits of the enzyme to the enzymatically active R state, and the inhibitor PEP, which shifts the four subunits to the inactive T state. The active site in the R state has a thousandfold higher affinity for the substrate F6P than does the active site of the T state, due to structural differences between the active sites in these two states. In the cell the activity of PFK is controlled through this allosteric mechanism by the relative concentrations of the two effectors. The inhibitor or negative effector is PEP, which is the product of an enzyme downstream of PFK in the glycolytic pathway. As the concentration of PEP increases it inhibits PFK and downregulates the pathway. This is the classic case of feedback inhibition.

Petitioner Merck, Ex. 1014, p. 118

Selected readings

General

Cohen, F.E., et al. Structural clues to prion replication. *Science* 264: 530–531, 1994.

Creighton, T.E. Up the kinetic pathway. *Nature* 356: 194–195, 1992.

Dobson, C.M. Finding the right fold. *Nature (Struct. Biol.)* 2: 513–517, 1995.

Fersht, A.R. Characterizing transition states in protein folding: an essential step in the puzzle. *Curr. Opin. Struct. Biol* 5: 79–84, 1994.

Finn, B.E., Forsen, S. The evolving model of calmodulin structure, function and activation. *Structure* 3: 7–11, 1995.

Freedman, R.B. The formation of protein disulfide bonds. *Curr. Opin. Struct. Biol.* 5: 85–91, 1995.

Goldsmith, E.J, Mottonen, J. Serpins: the uncut version. *Structure* 2: 241–244, 1994.

Hartl, F.U. Molecular chaperones in cellular protein folding. *Nature* 381: 571–580, 1996.

Lorimer, G.H. GroEL structure: a new chapter on assisted folding. *Structure* 2: 1125–1128, 1994.

Matthews, C.R. Pathways of protein folding. *Annu. Rev. Biochem.* 62: 653–683, 1993.

Miranker, A., Dobson, C.M. Collapse and cooperativity in protein folding. *Curr. Opin. Struct. Biol.* 6: 31–42, 1996.

Morgan, D.O. Principles of CDK regulation. *Nature* 374: 131–134, 1995.

Perutz, M. *Mechanisms of cooperativity and allosteric regulation in proteins*. Cambridge: Cambridge University Press, 1990.

Pines, J. Conformational change. *Nature* 376: 294–295, 1995.

Ptitsyn, O.B. Structures of folding intermediates. *Curr. Opin. Struct. Biol.* 5: 74–78, 1995.

Radzio-Andzelm, E.R., Lew, J., Taylor, S. Bound to activate: conformational consequences of cyclin binding to CDK2. *Structure* 3: 1135–1141, 1995.

Saibil, H.R. The lid that shapes the pot: structure and function of the chaperonin GroES. *Structure* 4: 1–4, 1996.

Weissman, J.S. All the roads lead to Rome? The multiple pathways of protein folding. *Chem. Biol.* 2: 255–260, 1995.

Specific structures

Babu, Y.S., et al. Three-dimensional structure of calmodulin. *Nature* 315: 37–40, 1985.

Baker, D., Sohl, J.L., Agard, D.A. A protein-folding reaction under kinetic control. *Nature* 356: 263–265, 1992.

Braig, K., et al. The crystal structure of the bacterial chaperonin GroEL at 2.8 Å. *Nature* 371: 578–586, 1994.

Brown, N.R., et al. The crystal structure of cyclin A. *Structure* 3: 1235–1247, 1995.

Carrell, R.W., et al. Biological implications of a 3 Å structure of dimeric antithrombin. *Structure* 2: 257–270, 1994.

Carrell, R.W., Evans, D.L., Stein, P.E. Mobile reactive centre of serpins and the control of thrombosis. *Nature* 353: 576–578, 1991.

Chen, S. et al. Location of a folding protein and shape changes in GroEL–GroES complexes imaged by cryo-electron microscopy. *Nature* 371: 261–264, 1994.

Creighton, T.E. The disulfide folding pathway of BPTI. *Science* 256: 111–114, 1992.

De Bondt, H.L., et al. Crystal structure of cyclin-dependent kinase 2. *Nature* 363: 595–602, 1993.

Dobson, C.M., Evans, P.A., Radford, S.E. Understanding protein folding: the lysozyme story so far. *Trends Biol. Sci.* 19: 31–37, 1994.

Fenton, W.A., et al. Residues in chaperonin GroEL required for polypeptide binding and release. *Nature* 371: 614–619, 1994.

Hunt, J.F., et al. The crystal structure of the GroES co-chaperonin at 2.8 Å resolution. *Nature* 379: 37–45, 1996.

Ikura, M., et al. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science* 256: 632–638, 1992.

Jeffrey, P.D., et al. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* 376: 313–320, 1995.

Kabsch, W., et al. Atomic structure of the actin:DNAse I complex. *Nature* 347: 37–44, 1990.

Kallen, J., et al. Structure of human cyclophilin and its binding site for cyclosporin A determined by x-ray crystallography and NMR spectroscopy. *Nature* 353: 276–279, 1991.

Loebermann, H., et al. Human α_1-proteinase inhibitor. *J. Mol. Biol.* 177: 531–556, 1984.

Martin, J.L., Bardwell, J.C.A., Kuriyan, J. Crystal structure of the DsbA protein required for disulphide bond formation *in vivo*. *Nature* 365: 464–468, 1993.

Mauguen, Y., et al. Molecular structure of a new family of ribonucleases. *Nature* 297: 3162–3164, 1982.

Mayhew, M., et al. Protein folding in the central cavity of the GroEL–GroES chaperonin complex. *Nature* 379: 420–426, 1996.

Meador, W.E., Means, A.R., Quiocho, F.A. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin–peptide complex. *Science* 257: 1251–1255, 1992.

Mottonen, J., et al. Structural basis of latency in plasminogen activator inhibitor-1. *Nature* 355: 270–273, 1992.

Schirmer, T., Evans, P.R. Structural basis of the allosteric behaviour of phosphofructokinase. *Nature* 343: 140–145, 1990.

Petitioner Merck, Ex. 1014, p. 119

Schreuder, H.A., et al. The intact and cleaved human antithrombin III complex as a model for serpin–proteinase interactions. *Nature (Struct. Biol.)* 1: 48–54, 1994.

Stein, P.E., et al. Crystal structure of ovalbumin as a model for the reactive centre of serpins. *Nature* 347: 99–102, 1990.

Takahashi, N., Hayano, T., Suzuki, M. Peptidyl-prolyl *cis–trans* isomerase is the cyclosporin A binding protein cyclophilin. *Nature* 337: 473–475, 1989.

Weissman, J.S., Kim, P.S. Kinetic role of non-native species in the folding of bovine pancreatic trypsin inhibitor. *Proc. Natl. Acad. Sci. USA* 89: 9900–9904, 1992.

Weissman, J.S., Kim, P.S. Re-examination of the folding of BPTI: predominance of native intermediates. *Science* 253: 1386–1393, 1992.

Yu, M.-H., Lee, K.N., Kim, J. The Z type variation of human α_1-antitrypsin causes a protein folding defect. *Nature (Struct. Biol.)* 2: 363–367, 1995.

Zhu, X., et al. Structural analysis of substrate binding by the molecular chaperone DnaK. *Science* 272: 1606–1614, 1996.

Petitioner Merck, Ex. 1014, p. 120