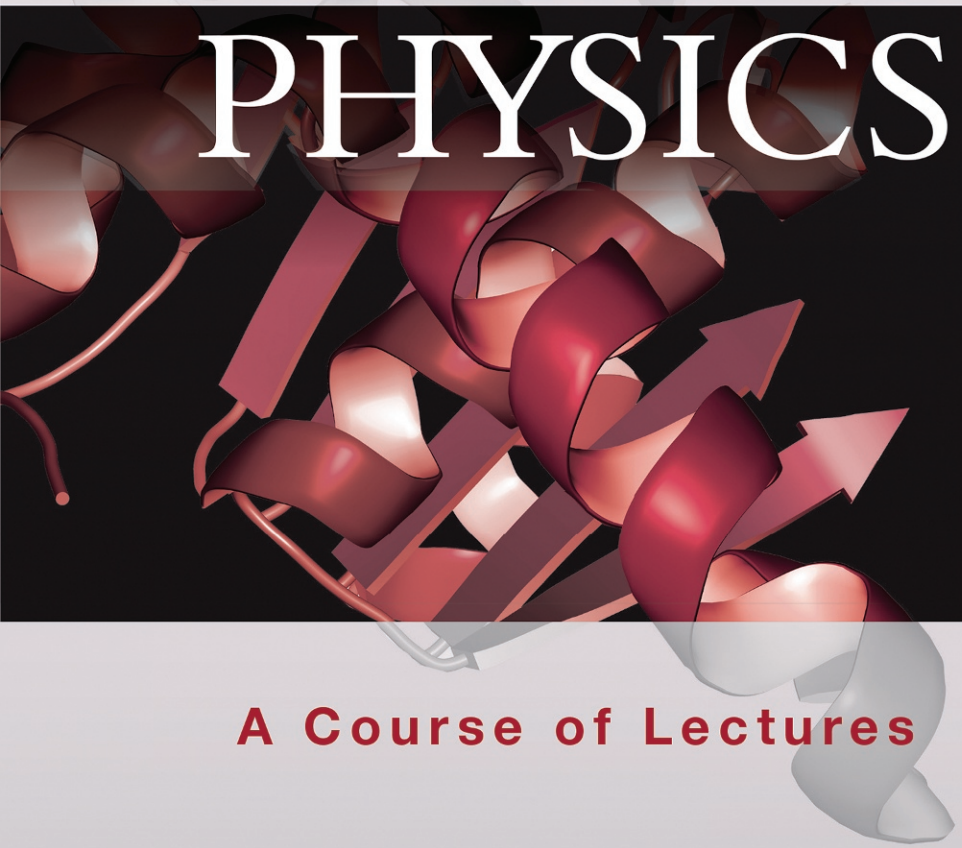
ELSEVIER

SECOND, UPDATED AND
EXTENDED EDITION

PROTEIN PHYSICS



A Course of Lectures

ALEXEI V. FINKELSTEIN
OLEG B. PTITSYN

Protein Physics

This page intentionally left blank

Protein Physics

Second, Updated and Extended Edition

Alexei V. Finkelstein
Oleg B. Ptitsyn*
Institute of Protein Research
Russian Academy of Sciences, Pushchino,
Moscow Region, Russian Federation

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

* Deceased

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

ISBN 978-0-12-809676-5

Publisher: John Fedor
Acquisition Editor: Anita Koch
Editorial Project Manager: Sarah Watson
Production Project Manager: Anitha Sivaraj
Cover Designer: Vicky Pearson Esser

Typeset by Spi Global, India

Contents

Foreword to the First English Edition

In June 1967, Oleg Ptitsyn became the Head of the Laboratory of Protein Physics at the new Institute of Protein Research at Pushchino. Three months later, he was joined by Alexei Finkelstein; first as a research student and then as a colleague. Their approach to the study of proteins was different from that common in the West, being strongly influenced by the Russian school of polymer physics. One of its most distinguished members, Michael Volkenstein, had been Ptitsyn's PhD supervisor. Together Ptitsyn and Finkelstein created, at Pushchino, one of the world's outstanding centers for the study of the physics and chemistry of proteins.

Certain areas of their work, particularly that on protein folding, have become well known in Europe, India and America, either directly through their papers or indirectly through the elaboration of their work by two of their former students: Eugene Shakhnovich, in America, and Alexei Murzin, in England. However, it became obvious when Finkelstein or Ptitsyn talked with colleagues, that the range of their work went far beyond what was commonly known in the West. We would find that fundamental questions that were our current concerns had already been considered by them and they had some elegant calculation that provided an answer. Usually, they had not got round to publishing this work. Now, to make their overall achievements available to more than just their friends and the students of Moscow University; Finkelstein has written this book on the basis of lectures he and Ptitsyn gave to these students.

In the breadth of its range, the rigor of its analysis and its intellectual coherence, this book is a *tour de force*. Of those concerned with the physics and chemistry of proteins, I doubt if there can be any, be they students or senior research workers, who will not find herein ideas, explanations and information that are new, useful and important.

Cyrus Chothia, FRC
MRC Laboratory of Molecular Biology
Cambridge, UK
2002

This page intentionally left blank

Preface

This book is devoted to protein *physics*, that is, to the overall topics of structure, self-organization and function of protein molecules. It is written as a course of lectures, which include a more-or-less conventional introduction to protein science (specifically, in the first and last parts of the book), as well as the story of in-depth protein science studies (in the middle of the book). Thus, the book is a kind of monograph embedded in the course of lectures, and it bridges the gap between introductory biophysical chemistry courses and research literature.

The course is based on lectures given by us (earlier by O.B.P. and later by A.V.F.), first at Moscow PhysTech Institute, then at Pushchino State University, and now at the Pushchino Branch of Lomonosov Moscow State University and at Biology and Bioengineering & Bioinformatics Departments of Lomonosov Moscow State University. Initially, our students were physicists, then mainly biologists with some chemists. That is why, by now, the lectures have not only been considerably up-dated (as research never stops) but also thoroughly revised to meet the requirements of the new audience.

Since what you will be reading has a form of a course of lectures, repetition is hardly avoidable (specifically, we have repeated some figures). Indeed, when delivering a lecture one cannot refer to "Figure 2 and Equation 3 of the previous lecture"; however, we have done our best to minimize repetition.

The following comments will help you to understand our approach and the way in which the material is presented:

On the "lecturer". All lectures are presented here in the way they are delivered, that is, from the first person—the lecturer.

On the "inner voice". The personality of the "lecturer" comprises both authors, O.B.P. and A.V.F., although this does not mean that no disagreement has ever occurred between us concerning the material presented! Moreover, sometimes each of us felt that the problem in question is disputable and subject to further studies. We made no attempt to smooth these disputes and contradictions over, so the "lecturer's" narration is sometimes interrupted or questioned by the "inner voice." Or the "inner voice" may simply articulate the frequently asked questions or elaborate the discussion.

Why "protein physics"? Because we are amazed to see how strongly biological evolution enhances, secures and makes evident the consequences of the physical principles underlying molecular interactions in the protein. It is

also striking how much our understanding of biological systems, proteins in particular, has developed through the application of physical methods. We see this from mass spectrometry and single-molecule techniques to electron or atomic force microscopy, X-ray crystallography and NMR studies of proteins. There is hardly any other area of contemporary science in which the traditional boundaries between disciplines and philosophies have been so clearly breached to such great profit.

On the physics and biology presented in these lectures. In the course of lecturing, we shall take the opportunity to present physical ideas, such as some elements of statistical physics and quantum mechanics. These ideas, to our minds, are absolutely necessary not only for an understanding of protein structure and function but also for general scientific culture, but usually a "normal" biology graduate has either completely forgotten or never known them. On the other hand, among a myriad of protein functions, we will discuss only those absolutely necessary to demonstrate the role of spatial structures of proteins in their biological, or rather biochemical, activities.

On "in vivo" and "in vitro" experiments. The terms "in vivo" and "in vitro", as applied to experiments, are often understood differently by physicists and biologists. Strictly speaking, between the pure "in vivo" and the pure "in vitro" there are a number of ambiguous intermediates. For example, protein folding in a cell-free system (with all its ribosomes, initiation factors, chaperones, crowding, etc.) is unequivocally an "in vivo" experiment in the physicist's view (for a physicist, "in vitro" would be mostly a separate protein in solution; even the cell-free system contains too many biological details). But for a biologist, this is undoubtedly an "in vitro" experiment (since "in vivo" is referred to a living and preferably intact organism). However, structural studies of a separate protein in an organism are hardly possible. Therefore, reasonable people compromise by making biologically significant "in vivo" events accessible for experimental "in vitro" studies.

On experiment, physical theory and calculations. Experiment provides the basic facts underlying all our ideas of phenomena, as well as a lot of refining details. However, the experimental methods used in protein physics are only briefly described in this book, with references to excellent textbooks that describe them in detail. The same, and to an even greater extent, concern chemical and biochemical methods (such as those used in purification of proteins or genetic engineering), which are just barely mentioned in the book.

Theory allows us to understand the essence and interrelation of the phenomena and is helpful in planning informative experiments. Some basic physical theories—in a simplified form, naturally—are included in our lectures, not only because they permit us to put in order and comprehend from a common standpoint the vast experimental material, but also because they are elegant. Besides, we believe that knowledge of basic physical theories and models is essential to human culture.

Calculations connect theory with experiment and verify key points of the theory. However, not everything that can be calculated must be calculated; for example, it is easier just to measure water (or protein) density than to calculate it from first principles. And not only is this easier but it yields a more accurate result, since a detailed calculation requires many parameters that can hardly be accurately estimated.

On physical models, rough estimates and computer (in silico) experiments. In these lectures, we will often discuss simple models, that is, those drastically simplified compared with reality, and use rough estimates. And we want you, after reading these lectures, to be able to make such estimates and use simple models of the events in question. The use of simple models and rough estimates may seem to be quite old-fashioned. Indeed, it is often believed that with the powerful computers available now, one can enter "all as it is in reality": water molecules, salt, coordinates of protein atoms, DNA, etc. fix the temperature, and obtain "the precise result." As a matter of fact, this is a Utopian picture. The calculation—we mean a detailed one (often made using so-called molecular dynamics)—will take days and cover only some nano- or microseconds of the protein's life, because you will have to follow the thermal motions of many thousands of interacting atoms. In any case, this calculation will not be absolutely accurate either, since all elementary interactions can be estimated only approximately. And the more detailed the description of a system is, the more elementary interactions are to be taken into account, and the more minor errors will find their way into the calculation (not to mention the increased computer time required). Eventually, you will obtain only a more-or-less precise estimate of the event instead of the desired absolutely accurate description of it—with days of a supercomputer's time spent. Meanwhile, what really interests you may be a simple quick estimate such as whether it is possible to introduce a charge into the protein at a particular site without a risk of protein structure explosion. That is why one of our goals is to teach you how to make such estimates. However, this does not mean that we will simply ignore computer experiments. Such experiments yield a lot of useful information. But the computer experiment is a real experiment (although in silico, not in vitro or in vivo). It involves highly complex systems and yields facts requiring further interpretation, which, in turn, demands simplified but clear models and theories.

On equations. We are aware that mathematical equations can be a difficulty for biologists, so we have done our best to refrain from using them, and only those really unavoidable have survived (some very useful but more complicated equations are used in the Appendices and Problem sections). Our advice is: when reading these lectures, "test words by equations." It is certainly easier for biologists to read words only, but they are often ambiguous, so word-verifying by equations and vice versa will help your understanding. To avoid going into insignificant (and unhelpful) detail, we shall often use

approximate calculations; therefore you will often see the symbols "≈" (approximately equal to) and "∼" (of the same order of magnitude as).
On references and figures. The references are included in the text, legends to figures and tables; the lists of references are positioned at the end of each Lecture or Appendix.

Protein structures are drawn using the programs MOLSCRIPT (Kraulis, P.J.,J. Appl. Cryst. 24, 946–950); WHAT IF (Vriend, G., J. Mol. Graphics 8, 52–56); RASMOL (Sayle, R., Milnerwhite, E.J., Trends Biochem. Sci. 24, 374–376); Insight II (Molecular Simulations Inc., 1998); ViewerLite (Accelrys Inc., 2001); and coordinates taken from the Protein Data Bank (initially described in: Bernstein, F.C., Koetzle, T.F., Meyer, E.F., Jr., Brice, M.D., Kennard, O., Shimanouchi, T., Tasumi, T., J. Mol. Biol. 112, 535–542). Many figures are adapted from the literature, with appropriate references and permissions. Most of other figures are purposely schematic.

On tastes. These lectures, beyond doubt, reflect our personal tastes and predilections and are focused on the essence of things and events rather than on a thorough description of their details. In the main, they contain physical problems and theories, while only the necessary minimum of experimental facts are given and experimental techniques are barely mentioned. (Specifically, almost nothing will be said in these lectures about the techniques of X-ray crystallography and NMR spectroscopy that have provided the bulk of our knowledge of protein structure; we tried to compensate this by reference to excellent textbooks on experimental methods.)

Therefore, these lectures are by no means a substitute for regular fact-rich biophysical and biochemical courses on proteins. When referring to specific proteins, we merely give the most important (from our viewpoint) instances; only absolutely necessary data are tabulated; all values are approximate, etc.

On small print. This is used for helpful but not essential excursus, additions and explanations.
On the personal note in these lectures. We shall take the opportunity of noting our own contributions to protein science and those by our co-workers and colleagues from the Institute of Protein Research, Russian Academy of Sciences. This will certainly introduce a "personal note" into the lectures and perhaps make them a bit more vivid.
What is the difference between 1st and 2nd English editions of Protein Physics? To my (A.V.F.) surprise, the current revision of the book showed no need of any correction of the initial text. I have made only a few cuts there. But of course I have updated the book to include new fascinating material. Specifically, this concerns novel information on intrinsically disordered proteins, amyloid aggregation, protein folding in vivo, protein motors, misfolding, chameleon proteins, advances in protein engineering and design and advances in modelling of protein folding. Also, the revised version includes a Problems section (with solutions).

Acknowledgements

We are grateful to all co-authors of our works that we have used in this book.

We are grateful to our colleagues from the Institute of Protein Research, Russian Academy of Sciences, and especially to former and current co-workers from our Laboratory of Protein Physics: this book would never have been written without scientific and friendly human contact with them.

We thank our university students from different countries, participants of various summer schools and readers of all previous (English, Russian, Chinese) editions of this book for their questions and comments that helped us to clarify (primarily, to ourselves) some points of this course and were helpful for better presentation of the lectures.

We are most grateful to A.B. Chetverin, A.S. Spirin, V.A. Kolb, A.V. Efimov, D.I. Kharakoz, V.E. Bychkova, Yu.V. Mitin, M.A. Roytberg, V.V. Velkov, I.G. Ptitsyna, G.I. Gitelson, D.S. Rykunov, G.V. Semisotnov, A.V. Skugaryev, A.K. Tsaturyan, C.Y. Bershitsky, D.N. Ivankov, A.Ya. Badretdinov, S.O Garbuzynskiy, O.V. Galzitskaya, and A. Li for reading the manuscript, discussions and most useful comments, and to A.A. Vedenov, A.M. Dykhne, M.D. Frank-Kamenetskii, V.N. Uversky, V.E. Finkelstein, A.A. Klimov, D.A. Klimov, F.K. Gioeva, A.M. Gutin, A.G. Murzin, E.I. Shakhnovich, I.M. Gelfand, B.A. Reva, E.S. Nadezhdina, E.N. Samatova-Baryshnikova, V.G. Sharapov, L.B. Pereyaslavets, C. Chothia, M. Levitt, M.S. Gelfand, A. Kister, A. Alexandrescu, C. Dobson, A. Fersht, G. Vriend, A. Tramontano, A. Lesk, K.W Plaxco and D. Baker for discussing various items touched upon in this book and supplying us with some additional materials or calculations.

Special thanks to D.S. Rykunov, G.A. Morozov, A.V. Skugaryev, D.N. Ivankov, M.Yu. Lobanov, N.Yu. Marchenko, N.S. Bogatyreva, M.G. Nikitina, V.G. Semisotnova I.V. Sokolovsky, A.A. Shumilin, M.G. Dashkevich, T.Yu. Salnikova, A.E. Zhumaev, A.A. Finkelstein and M.S. Vilner-Marmer for their valuable assistance in the preparation of the manuscript and figures, and to E.V. Serebrova for preparation of the English translation of the current and preceding editions of this book.

And, last but not least, we are grateful to the Russian Foundation for Basic Research, the "Molecular and Cell Biology" program of the Russian Academy of Sciences, the Moscow Grants, the Russian grants for Leading Scientific Schools and to the Howard Hughes Medical Institute (USA) for the financial support of our laboratory that enabled us to complete the previous editions of this book.

The preparation of the current edition is supported by a grant from the Russian Science Foundation.

To my deep sorrow, I (A.V.F.) had to complete the first edition and now to revise and update this new edition of the book alone: my teacher and co-author Oleg B. Ptitsyn passed away in 1999 … Therefore, I am completely responsible for any deficiencies in this book, and I am indeed indebted to Oleg B. Ptitsyn for his contribution to the composition and initial editing of the book.

Alexei V. Finkelstein, 2016

Part I

Introduction

This page intentionally left blank

Lecture 1

This lecture contains an introduction to the whole course—a brief overview of what is given (or omitted) in the following lectures. Therefore, unlike other lectures, this one is not supplied with specific references; instead, it contains a list of textbooks that may be recommended for additional reading.

Proteins are molecular machines, building blocks, and arms of a living cell. Their major and almost sole function is enzymatic catalysis of chemical conversions in and around the cell. In addition, regulatory proteins control gene expression, and receptor proteins (which sit in the lipid membrane) accept intercellular signals that are often transmitted by hormones, which are proteins as well. Immunoproteins and the similar histocompatibility proteins recognize and bind "foe" molecules as well as "friend" cells, thereby helping the latter to be properly accommodated in the organism. Structural proteins form microfilaments and microtubules, as well as fibrils, hair, silk, and other protective coverings; they reinforce membranes and maintain the structure of cells and tissues. Transfer proteins transfer (and storage ones store) other molecules. Proteins responsible for proton and electron transmembrane transfer provide for the entire bioenergetics, that is, light absorption, respiration, ATP production, etc. By ATP "firing" other proteins provide for mechanochemical activities—they work in muscles or move cell elements.

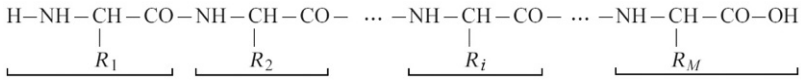
The enormous variety of protein functions is based on their high specificity for the molecules with which they interact, a relationship that resembles a key and lock (or rather, a somewhat flexible key and a somewhat flexible lock). This specific relationship demands a fairly rigid spatial structure of the protein—at least when the protein is "operating" (before and after that, some proteins are "natively unfolded"). That is why the biological functions of proteins (and other macromolecules of the utmost importance for life—DNA and RNA) are closely connected with the rigidity of their three-dimensional (3D) structures. Even a little damage to these structures, let alone their destruction, is often the reason for loss of, or dramatic changes in, protein activities.

A knowledge of the 3D structure of a protein is necessary to understand how it functions. Therefore, in these lectures, the physics of protein function will be discussed after protein structure, the nature of its stability and its ability to self-organize, that is, close to the end of this course.

Proteins are polymers; they are built up by amino acids that are linked into a peptide chain; this was discovered by E. Fischer as early as the beginning of the 20th century. In the early 1950s, F. Sanger showed that the sequence of amino acid residues (a "residue" is the portion of a free amino acid that remains after polymerization) is unique for each protein. The chain consists of a chemically regular backbone (main chain) from which various side chains

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00001-6

(R_1, R_2, …, R_i, …, R_M) project (later on, we will consider certain deviations from the backbone, —NH—CH—CO—, regularity):



The number M of residues in protein chains ranges from a few dozens to many thousands. This number is gene-encoded.

There are 20 main (and a couple of accessory) species of amino acid residues. Their position in the protein chain is gene-encoded, too. However, subsequent protein modifications may contribute to the variety of amino acids.

Also, some proteins bind various small molecules, serving as cofactors.

In an "operating" protein, the chain is folded in a strictly specified structure. In the late 1950s, Perutz and Kendrew solved the first protein spatial structures and demonstrated their highly intricate and unique nature. However, it is noteworthy that the strict specificity of the 3D structure of protein molecules was first shown (as it became clear later) back in the 1860s, by Hoppe-Zeiler who obtained hemoglobin crystals—in a crystal each atom occupies a unique place.

The question whether the structure of a protein is the same in a crystal (where protein structures had been first established) and in a solution had been discussed for many years (when only indirect data were available) until the virtual identity of these (apart from small fluctuations) was demonstrated by nuclear magnetic resonance (NMR) spectroscopy.

Proteins "live" under various environmental conditions, which leave an obvious mark on their structures. The less water there is around, the more valuable the hydrogen bonds are (which reinforce the regular, periodic 3D structures of the protein backbone) and the more regular the stable protein structure ought to be.

According to their "environmental conditions" and general structure, proteins can be roughly divided into three classes:

1. Fibrous proteins form vast, usually water-deficient aggregates; their structure is usually highly hydrogen-bonded, highly regular, and maintained mainly by interactions between various chains.
2. Membrane proteins reside in a water-deficient membrane environment (although they partly project into water). Their intramembrane portions are highly regular (like fibrous proteins) and highly hydrogen-bonded, but restricted in size by the membrane thickness.
3. Water-soluble (residing in water) globular proteins are less regular (especially small ones). Their structure is maintained by interactions of the chain with itself (where an important role is played by interactions between

hydrocarbon—"hydrophobic"—groups that are far apart in the sequence but adjacent in space) and sometimes by chain interactions with cofactors.

Finally, there are some, mostly small or hydrocarbon group-poor or charged group-rich polypeptides, which do not have an inherent fixed structure in physiological conditions by themselves but obtain it by interacting with other molecules. They are usually called "natively" (or "intrinsically") disordered (or unfolded) proteins.

The above classification is extremely rough. Some proteins may comprise a fibrous "tail" and a globular "head" (eg, myosin), and so on.

To date, we know many millions of protein sequences (they are deposited at special computer databanks, eg, Swiss-Prot) and hundreds of thousands of protein spatial structures (they are compiled at the Protein Data Bank, or simply PDB). What we know about 3D protein structures mostly concerns water-soluble globular proteins. The solved spatial structures of membrane and fibrous proteins are relatively few. The reason is simple: water-soluble proteins are easily isolated as separate molecules, and their structure is relatively easily established by X-ray crystallography and by NMR studies in solution. That is why, when speaking about "protein structure" and "protein structure formation" one often actually means regularities shown for water-soluble globular proteins only. This must be kept in mind when reading books and papers on proteins, including these lectures. Moreover, it must be kept in mind that, for the same experimental reason, contemporary protein physics is mainly physics of small proteins, while the physics of large proteins is only starting to develop.

Noncovalent interactions maintaining 3D protein architecture are much weaker than chemical bonds fixing a sequence of monomers (amino acids) in the protein chain. This sequence—it is called "the primary structure of a protein" (Fig. 1.1)—results from biochemical matrix synthesis according to a gene-coded "instruction."

Protein architectures, especially those of water-soluble globular proteins, are complex and of great diversity, unlike the universal double helix of DNA (the single-stranded RNAs appear to have an intermediate level of complexity). Nevertheless, certain "standard" motifs are detected in proteins as well, which will be discussed in detail in the last half of this course (note that the "standard" structures are, in fact, the same in all kingdoms of living matter).

In the first place, proteins have regular secondary structures, namely, the α-helices and β-sheets; α-helices are often represented by helical ribbons, and extended β-structural regions (which by sticking together form sheets) by arrows (see Fig. 1.1). Secondary structures are characterized by a regular periodic shape (conformation) of the main chain with side chains of a variety of conformations.

The packing of the secondary structures of one polypeptide chain into a globule is called the "tertiary structure," while several protein chains integrated into a "superglobule" form the "quaternary structure" of a protein. For instance

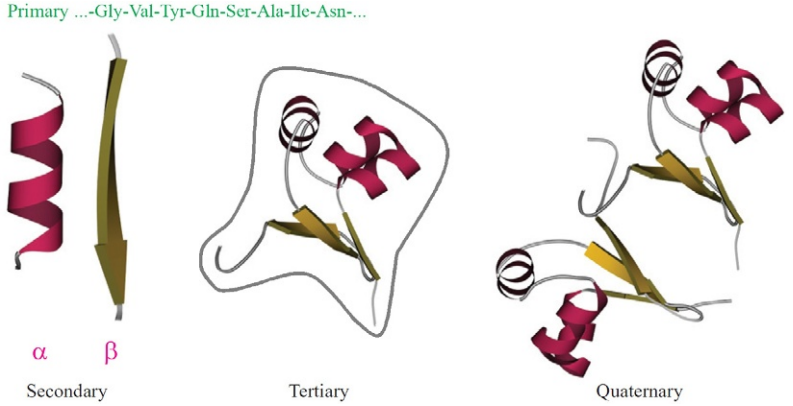


FIG. 1.1 Levels of protein structure organization: primary structure (amino acid sequence); regular secondary structures (α-helix and one strand of β-structure are shown); tertiary structure of a globule formed by one chain (the gray contour outlines the body of a dense globule); and quaternary structure of an oligomeric protein formed by several chains (here, dimeric *cro* repressor). This figure and most of the others show schemes of the backbone folds only; an all-atom presentation of a protein is given, eg, in Fig. 1.3.

(another example in addition to the dimeric *cro* repressor shown in Fig. 1.1), hemoglobin consists of two β- and two α-chains (this has *nothing to do* with α- and β-structures!). A quaternary structure formed by identical chains usually appears to be symmetrical (*cro* repressor and hemoglobin are no exception). Sometimes a quaternary structure comprises tens of protein chains. Specifically, virus coats can be regarded as such "superquaternary" structures (which are left unconsidered here).

Among tertiary structures, some can be distinguished as the most typical, and we will consider these later. They often envelop not the entire globular protein but only compact subglobules (so-called domains) within it (Fig. 1.2). A domain (like a small protein) usually consists of 100–200 amino acid residues, ie, of ∼2000 atoms. Its diameter is about 30–40 Å.

The in vivo formation of the native (ie, biologically active) tertiary structure occurs during biosynthesis or immediately after. However, it is noteworthy that a 3D structure can result other than from biosynthesis: around 1960, Anfinsen showed that it could also be yielded by "renaturation," ie, by in vitro refolding of a somehow unfolded protein chain and that the renaturation process goes spontaneously, unaided by the cellular machinery. This means that the spatial structure of a protein is determined by its amino acid sequence alone (provided water temperature, pH, and ionic strength are favorable), ie, the protein structure is capable of *self*-organizing. Later on we will consider certain exceptions.

Inner voice: Strictly speaking, this has been shown mainly for relatively small (up to 200–300 amino acid residues) water-soluble *globular* proteins.

FIG. 1.2 The domain structure of a large protein is similar to the quaternary structure built up by small proteins. The only difference is that in large proteins, the compact subglobules (domains) pertain to the same chain, while the quaternary structure comprises several chains.

> Concerning larger proteins, especially those of higher organisms, it is not that simple: far from all of them refold spontaneously…
> *Lecturer*: Thanks for the refinement. Yes, it is not that simple with such proteins. Part of the reason is aggregation, part is posttranslational modification, especially when it comes to higher organisms, eukaryotes. Still less is known about spontaneous self-organization of membrane and fibrous proteins. In some proteins of this kind it happens, but mostly they do not refold. Therefore, let us agree right away that when speaking about protein physics, protein structure and its formation, I will actually discuss (if not specified otherwise) relatively small, single-domain globular proteins. This convention is quite common for biophysical literature, but the fact that it is insufficiently articulated causes frequent misunderstandings.

Anfinsen's experiments provided fundamental detachment of the physical process, that is, the spatial structure organization, from biochemical synthesis of the protein chain. They made clear that the structure of protein is determined by its own amino acid sequence alone and is not imposed by cellular machinery. It seems that the main task of this machinery is to protect the folding protein from unwanted contacts (among these are also contacts between remote regions of a very large protein chain), since in vivo folding occurs in a cellular soup with a vast variety of molecules to stick to. But in diluted in vitro solution, a protein, at least a small one, folds spontaneously by itself.

Strictly speaking, proteins are capable of spontaneous refolding provided they have undergone no strong posttranslational modification, ie, if their chemical structure has not been strongly affected after biosynthesis and initial folding. For example, insulin (which loses half of its chain after its in vivo folding has been accomplished) is unable to refold.

Posttranslational modifications (which are hardly considered in this book) are of a great variety. As a rule, chemical modifications are provided by special

enzymes rather than "self-organized" within protein. First of all, I should mention cleavage of the protein chain (proteolysis: it often assists conversion of zymogen, an inactive proenzyme, into the active enzyme; besides, it often divides a huge "polyprotein" chain into many separate globules). The cleavage is sometimes accompanied by excision of protein chain fragments (eg, when deriving insulin from proinsulin); by the way—the excised fragments are sometimes used as separate hormones. Also, one can observe modification of chain termini, acetylation, glycosylation, lipid binding to certain points of the chain, phosphorylation of certain side groups, and so on. Even "splicing" of protein chains (spontaneous excising of a chain fragment and sticking the loose ends together) has been reported. Spontaneous cyclization of protein chains or their certain portions is also occasionally observed.

Particular attention has to be given to formation of S—S bonds between sulfur-containing Cys residues: "proper" S—S bonds are capable (under favorable in vitro conditions) of spontaneous self-formation, although in vivo their formation is catalyzed by a special enzyme, disulfide isomerase. S—S-bonding is typical mostly of secreted proteins (there is no oxygen in the cell, and consequently, no favorable oxidizing potential for S—S-bonding). The S—S bonds, if properly paused, are not at all harmful but rather useful for protein renaturation.

In contrast, "improper" S—S bonds prevent protein renaturation. That is why, by the way, a boiled egg does not "unboil back" as the temperature decreases (although I should say that, in 2015, the "Improbable Nobel Prize" was awarded, see http://www.improbable.com/ig/winners/, for inventing a chemical recipe on how to partially unboil back an egg). The reason is that high temperature not only denatures the egg's proteins but also initiates formation of additional S—S bonds between them (like in a gum). As the temperature is decreased, these new chemical bonds persist, thereby not allowing the egg's proteins to regain their initial (native) state.

Thus, the amino acid sequence of a protein determines its native spatial structure, and this structure, in turn, determines its function, ie, with whom this protein interacts and what it does.

Here, I have to make some additional comments:

First, Fig. 1.1 apparently shows that there is ample empty space in the interior of the protein, and can create an impression that the protein is "soft." In fact, this is not true. Protein is "hard": its chain is packed tightly, atoms against atoms (Fig. 1.3A). However, the space-filling representation is inconvenient for studying protein anatomy, its skeleton, its interior; these can be seen using the wire model or some of the above-presented schemes with "transparent" atoms and a clear pathway of the protein chain (see Fig. 1.3B and especially Figs. 1.1 and 1.2 where the side chain atoms are stripped off and secondary structure elements stand out).

The space-filling model (Fig. 1.3A) gives no idea even of the polymeric nature of protein; it only shows the surface of a globule looking potato-like. However, this model is useful for studying protein function, since it is the physicochemical and geometrical properties of the globule surface, the "potato

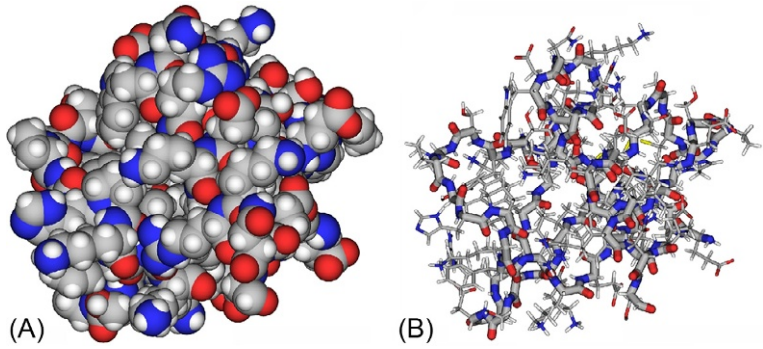


FIG. 1.3 Space-filling representation of a protein globule (A) and its wire model (B). In the wire model, side chains are shown as thin lines, the backbone as the thick line. Atoms are shown in colors: gray, C; white, H; red, O; blue, N.

skin," that determine the specificity of the protein activity, whereas the protein skeleton is responsible for the creation and maintenance of this surface.

Second, apart from the polypeptide chain, proteins often contain *cofactors* (Fig. 1.4), such as small molecules, ions, sugars, nucleotides, fragments of nucleic acids, etc. These nonpeptide molecules are involved in protein functioning and sometimes in the formation of protein structure as well. The cofactors

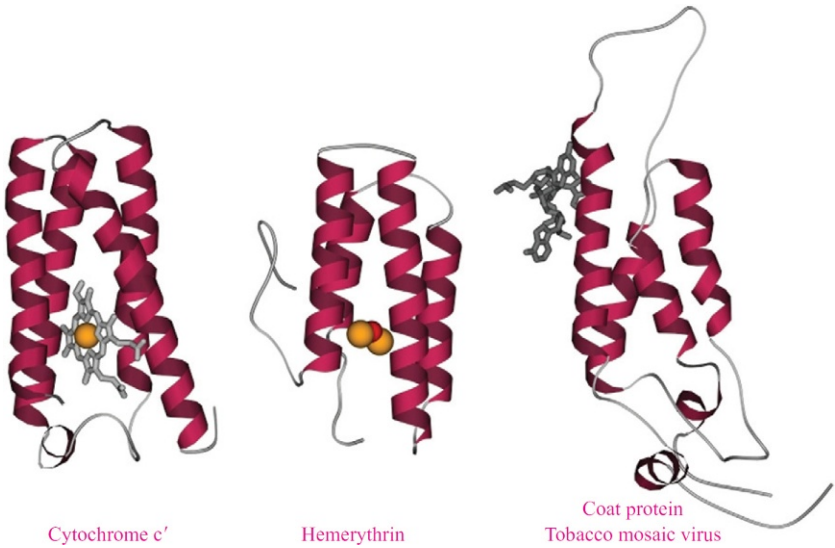


FIG. 1.4 Three α-helical proteins similar in overall architecture (comprising four α-helices each), but different in function: cytochrome *c′*, hemerythrin, and coat protein of Tobacco mosaic virus. Protein chain is shown as a ribbon; cofactors are shown as follows: wire models, heme (in cytochrome), and RNA fragment (in viral coat protein); orange balls, iron ions (in cytochrome heme and hemerythrin); a red ball, iron-bound oxygen (in hemerythrin).

can be linked by chemical bonds or just packed in cavities in a protein globule. Also, many water molecules (not shown in Fig. 1.4) are usually tightly bound to the protein surface.

Third, a solid protein ("aperiodic crystal" in Schrödinger's wording) behaves exactly like a crystal under varying conditions (eg, at increasing temperature), that is, for some time it "stands firm" and then melts abruptly, unlike glass, which loses its shape and hardness gradually. This fundamental feature of proteins is closely allied to their functional reliability: as with a light bulb, proteins become inoperative in the "all-or-none" manner, not gradually (otherwise their action would be unreliable, eg, cause low specificity, etc. We will discuss this later on).

Finally, as concerns hardness, we have to distinguish between relatively small single-domain proteins that are really hard (they consist of one compact globule) and larger proteins that have either a multidomain (Fig. 1.2) or quaternary (Fig. 1.1) structure. The component subglobules of larger proteins can somewhat move about one another.

In addition, like a solid body, all globules can become deformed (but not completely reorganized) in the course of protein functioning.

Proteins with similar interior organization (anatomy) usually have the same related function.

For example, many (though not all) cytochromes look like the one shown in Fig. 1.4, and many (though again not all) serine proteases from different species (from bacteria to vertebrates) look like chymotrypsin, as shown in Fig. 1.5.

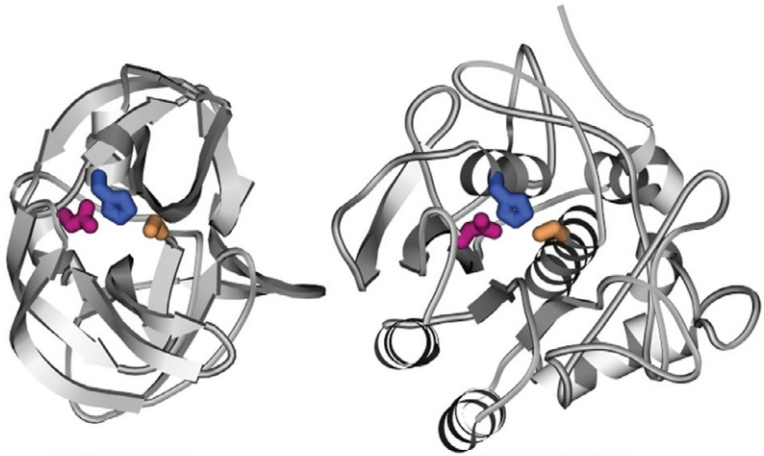


FIG. 1.5 Two proteins structurally different but almost identical in function (serine proteases): chymotrypsin, formed by β-structure, and subtilisin, formed by α-helices (some of which pertain to the active site) along with β-structure. In spite of drastically different chain folds, their catalytic sites comprise the same residues similarly positioned in space (but not in the amino acid sequence): Ser195 (orange), His57 (blue), and Asp102 (crimson) in chymotrypsin and Ser221 (orange), His64 (blue), and Asp32 (crimson) in subtilisin.

But sometimes very similar spatial structures may provide completely different functions. For example, cytochrome, one of the three proteins with similar spatial structures shown in Fig. 1.4, binds an electron, while hemerythrin, another protein of this shape, binds oxygen (these functions are somewhat alike, since they both are involved in the chain of oxidation reactions) and viral coat protein associates with much larger molecules, such as RNA and other coat proteins, and has nothing to do with oxidation.

We have already said that the structure of a protein determines its function. Is the reverse true? That is, does the function of a protein determine its structure?

Although some particular correlations of this kind have been reported, in general the influence of function on the structure has been detected mainly at a "rough" structural level, that is, the level connected with the "environmental conditions" of protein functions (eg, proteins controlling the structural function, like those building up hair or fibrils, are mostly fibrous proteins, receptors are membrane proteins, etc.). But most frequently, we see no influence of function on protein anatomy and architecture. For example, two serine proteases, chymotrypsin and subtilisin, have the same catalytic function and even similar specificity, whereas their interior organizations have nothing in common (see Fig. 1.5; their similarity is no greater than that between a seal and a diving beetle: only the proteins' "flippers," ie, their active sites including half a dozen of amino acid residues (from a couple of hundreds forming the globules), are structurally alike, while in every other respect they are completely different). Moreover, there exist structurally different active sites performing the same work (eg, those of serine proteases and metal proteases).

These, and many other examples, show that the function of a protein does not determine its 3D structure.

But, while saying this, account must be taken of size.

If the treated molecule (the molecule with which the protein interacts) is large, then a large portion of the protein may be involved in the interaction, and hence nearly the entire architecture of the protein is important for its functioning.

If the protein-treated molecule is small (which is more common, for enzymes in particular), then it is minor details of a small fraction of the protein surface that determine its function, while the rest of its "body" is responsible for fixing these crucial details. Hence, the main task of the bulk of the protein chain is to be hard and provide a solid foundation for the active site.

> *Inner voice*: The nontrivial and piquant facts that one and the same function is performed by proteins of utterly different architectures, and different functions by architecturally similar proteins, should not overshadow the absolutely correct commonplace that architecturally close proteins are often homologous (genetically related) and have identical or similar functions…
> *Lecturer*: This is true—but trivial. What I wanted to emphasize is the idea, important for protein physics (and protein engineering), that active sites may depend only slightly on the arrangement of the remaining protein body.
>
> And the common feature of "the native protein body" is its hardness, since there is no other way to provide active-site specificity.

In time, we will consider the structures of proteins, their ability to self-organize and the reason for their hardness; we will discuss their functions and other aspects of interest for a biologist; but first we have to study amino acid residues and their elementary interactions with one another and the environment, as well as secondary structures of proteins that form their frameworks. These will be the subjects of the next several lectures.

RECOMMENDED ADDITIONAL READING

Below I give a list of books that can be useful when reading these lectures.

This list does not include basic books on physics, chemistry, and mathematics for undergraduates. Also, it comprises no monographs, advanced books and collections dedicated to experimental and computational techniques and other special aspects of protein science, such as chemistry of protein chains and their chemical modifications; gene manipulations; enzymology; mathematical backgrounds of bioinformatics or proteomics; and other subjects that are related, but not directly related to these lectures.

Biochemistry and molecular biology textbooks:

Nelson, P., 2013. Biological Physics: With New Art by David Goodsell. W.H. Freeman & Co., New York.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co., New York.

Stryer, L., 1995. Biochemistry, fourth ed. W.H. Freeman & Co., New York.

Lehninger, A.L., Nelson, D.L., Cox, M.M., 1993. Principles of Biochemistry, second ed. Worth Publishers, New York.

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman & Co., New York.

Books on protein physics and physical chemistry:

Petsko, G.A., Ringe, D., 2003. Protein Structure and Function. Sinauer Associates, Sunderland, MA.

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman & Co., New York.

Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York.

Tanford, C., 1980. The Hydrophobic Effect, second ed. Wiley-Interscience, New York.

Books on physics applied to biomolecules:

Frauenfelder, H., 2010. The Physics of Proteins. An Introduction to Biological Physics and Molecular Biophysics. Springer, New York.

Dill, K.A., Bromberg, S., 2010. Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience, second ed. Garland Science, New York.
Grosberg, A.Yu., Khokhlov, A.R., 1994. Statistical Physics of Macromolecules. American Institute of Physics, New York.
Volkenstein, M.V., 1977. Molecular Biophysics. Academic Press, New York.

Books on protein structures and bioinformatics:

Lesk, A., 2010. Introduction to Protein Science: Architecture, Function, and Genomics, second ed. Oxford University Press, Oxford, NY.
Tompa, P., 2010. Structure and Function of Intrinsically Disordered Proteins. Chapman & Hall/CRC Press, Taylor & Francis Group, Boca Raton, FL.
Lesk, A., 2001. Introduction to Protein Architectures. Oxford University Press, Oxford, NY.
Perutz, M.F., 1992. Protein Structure. W.H. Freeman & Co., New York.
Branden, C., Tooze, J., 1991, 1999. Introduction to Protein Structure. Garland Science, New York.

Books on protein folding, function and engineering:

Nölting, B., 2010. Protein Folding Kinetics: Biophysical Methods. Springer, New York.
Howard, J., 2000. Mechanics of Motor Proteins and the Cytoskeleton. Sinauer Associates, Sunderland, MA.
Fersht, A., 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W.H. Freeman & Co., New York.
Fersht, A., 1985. Enzyme Structure and Mechanism, second ed. W.H. Freeman & Co., New York.

Books on experimental and computational methods in protein science:

Schlick, T., 2010. Molecular Modeling and Simulation: An Interdisciplinary Guide, second ed. Springer, New York.
Lesk, A.M., 2008. Introduction to Bioinformatics, third ed. Oxford University Press, Oxford.
Serdyuk, I.N., Zaccai, N.R., Zaccai, J., 2007. Methods in Molecular Biophysics: Structure, Dynamics, Function. Cambridge University Press, Cambridge, NY.
Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge.

This page intentionally left blank

Part II

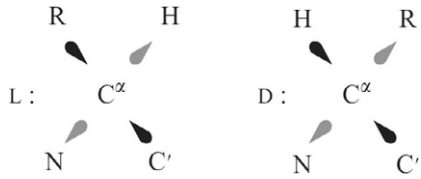
Elementary Interactions in and Around Proteins

This page intentionally left blank

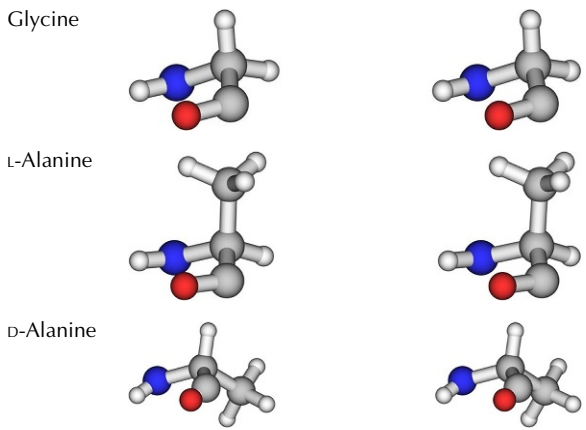
Lecture 2

The protein (polypeptide) chain consists of the main chain and side groups of amino acid residues (see Fig. 2.1) (Cantor and Schimmel, 1980; Schulz and Schirmer, 1979 & 2013; Nelson and Cox, 2012).

Amino acids that build up polypeptide chains (Figs. 2.1 and 2.2) can have either the L or the D steric form. The forms L and D are mirror-symmetric: the massive residue side chain (R) and the H-atom, both positioned at the α-carbon (C^{α}) of the amino acid, exchange places in these forms (arrows show atoms above and below the plane of the figure):



Glycine (Gly), with only a hydrogen atom as a side chain, shows no difference between the L- or D-form; for all other amino acids, the L- and D-forms have different shapes, as seen in the cross stereo drawings given here, for L-alanine and D-alanine:



(See How to use stereo drawings, in Appendix E.)

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00002-8

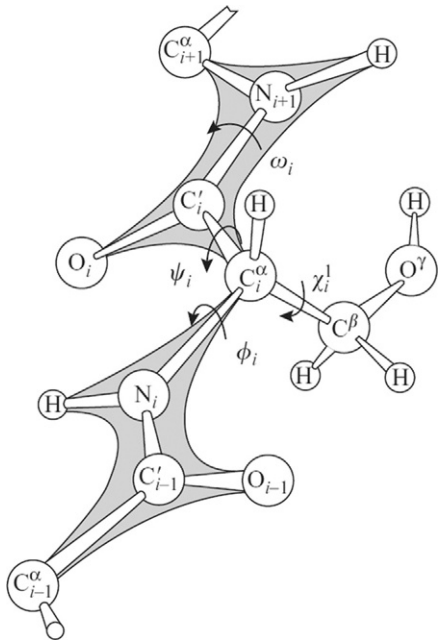
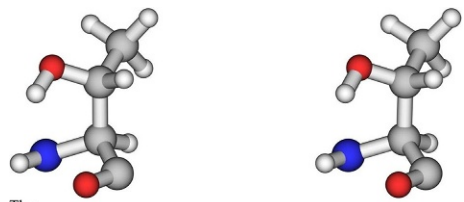


FIG. 2.1 Diagram showing a polypeptide chain with a side group (here: side group of Serine (Ser); '*i*' is its number in the chain). The peptide units are outlined. The main-chain angles of rotation (φ, ψ, ω) and that of the side chain (χ^1) are presented. Arrows show the direction of rotation of the part of the chain closest to the viewer about its remote part that increases the rotation angle. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979 & 2013. Principles of Protein Structure. Springer, New York, Chapter 2, with permission.)*

Protein chains are built up only from L-residues. Only these are gene-coded. D-residues (sometimes observed in peptides) are not encoded during matrix protein synthesis but made by special enzymes. Spontaneous racemization (L ↔ D transition) is not observed in proteins. It never occurs during biosynthesis but often accompanies chemical synthesis of peptides, and then its elimination is highly laborious.

All side chains have no L- and D-forms, but for Ile and Thr, where C^{γ} atoms have covalent bonds to four non-equal neighbours:



Only one of these forms (shown in the stereo drawing as an example) is present in proteins.

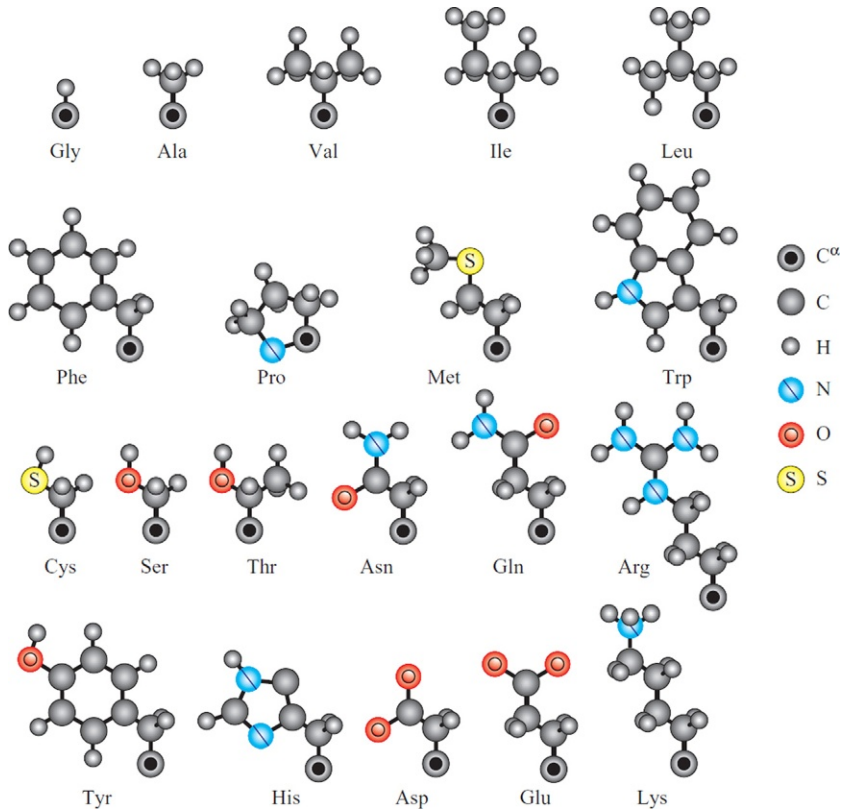
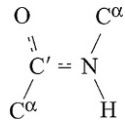


FIG. 2.2 The side chains of 20 standard amino acid residues (projecting from the main-chain C^{α} atoms). Atoms forming the amino acids are shown on the right. Apart from the 20 standard amino acid residues shown here, there are some rare non-standard ones, which are either produced by modification of a standard amino acid or (like selenocysteine) are coded by a RNA codon positioned in a special RNA context (Creighton, 1993).

In the protein chain, amino acids are linked by *peptide* bonds between C′- and N-atoms (Fig. 2.1). An important role is played by the planar rigid structure of the whole *peptide unit*:



Its planar character is provided by the so-called sp^2-hybridization of electrons of the N- and C′-atoms. 'Hybridization' of electron orbitals is a purely quantum effect. sp^2-Hybridization turns one spherical s-orbital and two '8-shaped'

p-orbitals into three extended (from the nucleus) sp^2-orbitals. These three orbitals involve the atom in three covalent bonds pertaining to one plane (-●<). A covalent bond is created by a 'delocalized' electron cloud covering the bound atoms.

The peptide group is rigid owing to an additional bond formed by p-electrons from the N- and C′-atoms uninvolved in sp^2-hybridization. These electrons of N-, C′-, and O-atoms also bond and 'delocalize', thereby creating an electron cloud that envelops all these atoms (that is why the bonds C=–N and C=–O are drawn as 'partial' double bonds, =–). And since the '8-shaped' p-orbitals are perpendicular to the plane of the sp^2-orbitals (-●<), the additional covalent bond of these perpendicular p-orbitals

blocks the rotation around the C′–N bond.

I would like to remind you that chemical bonds are caused mainly by 'delocalization' of electrons, ie, their permanent transition from one atom to another. This follows from Heisenberg's *Principle of Uncertainty*:

$$\Delta p \Delta x \sim \hbar \tag{2.1}$$

(more rigorously: $\Delta p \Delta x \geq \hbar/2$; Landau and Lifshitz, 1977).

Here Δp is the uncertainty in impulse ($p = mv$) of the particle, Δx is the uncertainty in its coordinate, and the reduced Planck's constant $\hbar \equiv h/2\pi$, where h is Planck's constant. Since the direction of electron movement within the atom is unpredictable, $\Delta p \sim |p| = mv$, where v is the velocity and m is the mass of the particle. Hence,

$$|v| \sim \frac{\hbar}{m\Delta x} \tag{2.2}$$

At the same time, the kinetic energy of the particle $E = mv^2/2$, ie, (neglecting insignificant '2'):

$$E \sim \frac{\hbar^2}{m\Delta x^2} \tag{2.3}$$

Hence, owing to the delocalization, the energy of the particle decreases with increasing Δx, and thereby the particle adopts a more stable state. As seen, light particles (electrons) are those mostly affected. This is how electron delocalization causes chemical bonding (Pauling, 1970).

The length of a chemical bond is close to the van der Waals radius of atoms, ie, it amounts to 1–2 Å (to be more exact, it is 1 Å for C–H, N–H and O–H bonds, about 1.2–1.3 Å for C=–O, C=–O, C=–N and C=–C, 1.5 Å for C–C and about 1.8 Å for S–S).

Typical values of covalent angles are approximately 120° and 109°. The former are at sp^2-hybridized atoms like –C′<, –N<, where three covalent bonds are directed from the center to the apexes of a planar triangle, and the latter are at sp^3-hybridized atoms, like >C^{α}<, where four bonds are directed from the center to the apexes of a tetrahedron, as well as at O< or S< atoms having two bonds each (Pauling, 1970; Schulz and Schirmer, 1979 & 2013).

Now let us consider typical values of vibrations, ie, thermal vibrations of covalent bonds and angles. These can contribute to the flexibility of the protein chain.

Vibrational frequencies manifest themselves in the infrared (IR) spectra of proteins. Typical frequencies are as follows (Schulz and Schirmer, 1979 & 2013): $\nu \sim 7 \times 10^{13}$ s^{-1} for vibrations of the H atom (eg, in the bond C–H; the corresponding IR light wavelength $\lambda = c/\nu \sim 5$ μm, where c is the speed of light, 300,000 km s^{-1}). For vibrations of 'heavy' atoms and groups (eg, in the bond CH_3–CH_3), $\nu \sim 2 \times 10^{13}$ s^{-1} (then $\lambda = c/\nu \sim 15$ μm).

Are these vibrations excited at room temperature?

To answer this question, we have to compare heat energy per degree of freedom ('heat quantum', kT) with vibration energy. Let us estimate kT at 'normal' temperature. Here, T is absolute temperature in Kelvin ($T = 300$ K at 27 °C, ie, at about 'room' temperature; K denotes Kelvin), and k (sometimes written as k_B) is the Boltzmann constant (equal to ≈2 cal mol^{-1} K^{-1}, or 0.33×10^{-23} cal K^{-1} per particle, since one mole contains 6×10^{23} particles). Hence, at room temperature the 'heat quantum' $kT = 600$ cal mol^{-1}, or (600 cal)/(6×10^{23} particles), ie, 10^{-21} calories per particle.

The frequency ν_T corresponding to this heat quantum can be derived from the well-known equation $kT = h\nu_T$ (where Planck's constant $h \equiv 2\pi\hbar = 6.6 \times 10^{-34}$ J s $\equiv 1.6 \times 10^{-34}$ cal s; let me remind you that 1 calorie is equal to 4.2 joules, J). So, the characteristic frequency of thermal motions, ν_T, is equal to 7×10^{12} s^{-1} at $T = 300$ K, ie, at 27 °C.

The 'heat quantum' cannot induce vibrations with a higher frequency than its own (Pauling, 1970; Schulz and Schirmer, 1979).

Thus, at room temperature, covalent bonds are 'hard' and do not vibrate: their vibrational frequency $\nu \sim 2 \times 10^{13} - 7 \times 10^{13}$ s^{-1}, ie, an order of magnitude higher than $\nu_T = 7 \times 10^{12}$ s^{-1}.

However, vibrations of covalent bonds can be induced by IR light; this underlies the importance of IR spectroscopy of proteins (Creighton, 1993). It is IR spectroscopy that provides information about the vibrations of atoms, covalent bonds and covalent angles. The properties of these vibrations are first derived from experiments on small molecules and then used for protein investigations.

Covalent angles are less rigid than bond lengths, and therefore, they vibrate at room temperature; their vibrational frequency ranges from 10^{12} to 10^{13} s^{-1}. However, their typical amplitude amounts only to 5° (Schulz and Schirmer, 1979 & 2013).

Thus, covalent bond vibrations do not contribute to the flexibility of protein chain, and the contribution of vibrations of covalent angles is minor.

In fact, this flexibility (which implies the ability to fold into α-helices and globules) is provided by rotation (although not a completely free rotation—see later) *around* the covalent bonds. That is why the chain structure (configuration) is often described simply in terms of angles of rotation around covalent bonds—then it is called the 'conformation'. It should be noted, that the terms 'structure', 'configuration' and 'conformation' are often used as synonyms.

The relative position of atomic groups linked by a covalent bond is described by a dihedral angle (that is formed by two planes: in Fig. 2.3, one includes points **a**, **O′**, **O″** and the other **O′**, **O″**, **c**).

Figs. 2.1 and 2.3 illustrate the measurement of this angle. The measurement is made as described in school trigonometry, if we assume that the covalent bond closest to the viewer is the 'rotating arrow', the far bond is the 'coordinate axis', and the central one is the 'rotation axis'. As in trigonometry, the arrow's turn in a counter-clockwise direction increases the angle of rotation, while its clockwise movement decreases this angle.

The major information on rotation of atomic groups around covalent bonds is also provided by IR spectroscopy; again, I will discuss only the basic results obtained.

Fig. 2.4 illustrates the typical variation of energy in the case of rotation around the bond between two sp^3-hybridized atoms ($H_3C–CH_3$ and $CH_2C–CH_2C$ serve as examples). These bonds are typical of aliphatic side chains. Side-chain angles of rotation are called χ ('chi') angles (Fig. 2.1). The maximums of such triple (in accordance with symmetry of rotation about the $sp^3–sp^3$ bond) potentials (ie, potentials that have three maximums and three minimums within 360°) correspond to three eclipsed conformations (0°, 120° and 240°) that result in approach (and repulsion) of the electron clouds. The repulsion occurs because these electrons have been already involved in covalent bonding.

The resultant potential barriers of rotation around $H_3C–CH_3$ amount to about 3 kcal mol^{-1}, and the typical range of thermal fluctuations about these minimums (ie, deviations accompanied by energy increasing by kT) is 15–20°.

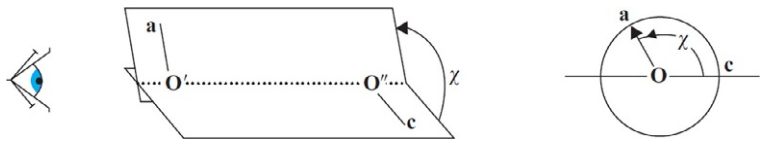


FIG. 2.3 Measurement of the dihedral angle (angle of rotation) shown in both axial (right) and transverse (left) views. The central bond **O′**–**O″** serves as the 'rotation axis', the covalent bond **O″**– **c** serves as the 'coordinate circle axis', and the closest (to the viewer) bond **O′**–**a** serves as the 'arrow on the coordinate circle' in trigonometry. When the dihedral angle is measured in the chain, the atoms **a** and **c** belong to the heaviest atomic groups attached to the **O′**–**O″** bond. (See IUPAC-IUB 1970, for more, sometimes rather tricky details of the dihedral angle nomenclature.)

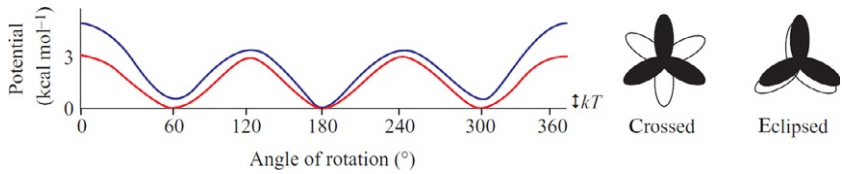


FIG. 2.4 Typical (see Halgren, 1995; Levitt et al., 1995; Jorgensen et al., 1996; Wang et al., 2004) potential of rotation around a single bond between two sp3-hybridized atoms: around $H_3C–CH_3$ (red curve) and $CH_2C–CH_2C$ (blue curve). The major energetic effect results from repulsion of electron clouds that is at a maximum in the 'shaded' conformations (at 0°, 120° and 240°) and at a minimum in the 'crossed' ones (at 60°, 180° and 300°). Repulsion of small H-atoms is negligible. However, repulsion of heavy C-atoms surrounded by large electron clouds occurring around 0 (in the chain C–C–C–C) yields an additional energetic effect that distinguishes rotation around the $(C–CH_2)–(CH_2–C)$ bond from that around the $H_3C–CH_3$ bond. For comparison, ↕ shows the magnitude of the 'heat quantum' kT.

When more massive atoms are sp^3-bonded instead of some H atoms, repulsion of these contributes to the barrier in the region where they become too close to one another. This is exemplified (see Fig. 2.4) by the rotation around the central bond in $(C–CH_2)–(CH_2–C)$.

Fig. 2.5 illustrates the typical variation of energy in the case of rotation around a peptide bond between two sp^2-hybridized atoms (C′ and N). The angle of rotation around this bond is denoted as ω (Fig. 2.1). The potential is double (ie, it has two maximums and two minimums within 360°) in accordance with the symmetry of rotation about the $sp^2–sp^2$ bond. The potential barriers are high owing to the involvement of additional p-electrons in the peptide bond (as discussed at the beginning of this lecture). The potential minimums are at 0° and 180° (where the p-orbitals pulling together C′ and N atoms are at their closest), and its maximums are at 90° and 270° (where these p-orbitals are farthest apart and, hence, least connected with one another). High barriers mean

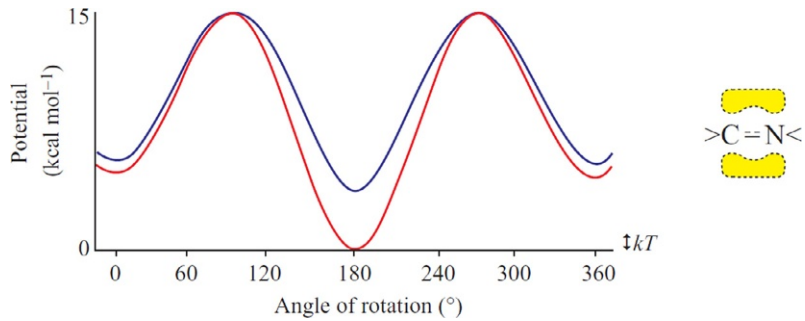
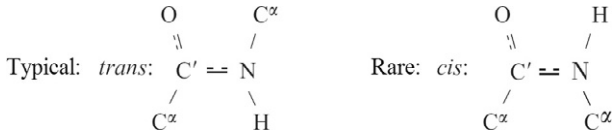


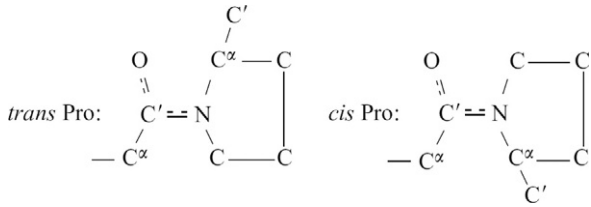
FIG. 2.5 Typical potential of rotation around a peptide bond between two sp^2-hybridized atoms (C′ and N); p-orbital-bonding at 0° (or 180°) is shown in yellow on the right. For all (except pre-proline) peptide bonds, their energy (red curve) is higher at 0° than at 180° owing to the repulsion between massive C^α-atoms at 0°. This difference in energy is small for the peptide bond preceding Pro (blue curve): Pro has not one but two C-atoms bonded to the N-atom. (See the text and the diagram illustrating the structural formula of proline.)

that the typical range of thermal fluctuations of the angle of rotation around such bonds is small (5–10°).

It is noteworthy that repulsion of massive C^{α}-atoms makes the *cis*-conformation ($\omega = 0°$) rather unfavourable energetically; therefore, in proteins, almost all peptide groups are in the *trans*-conformation ($\omega = 180°$).



An exception is the proline-preceding peptide bond. Pro is an *imino* but not an *amino* acid: its N atom has not two but three similar massive radicals (C′–, $-C^{\alpha}HC_2$ and $-CH_2C$; see Fig. 2.2), and therefore its *trans*-conformation has only a minor advantage as compared with the *cis* one.



In both globular and unfolded (ie, fluctuating, lacking a fixed structure) peptides, there are about 90% of *trans*- and 10% of *cis*-prolines (Branden and Tooze, 1999).

I would like to draw your attention to this regularity—the more favourable some detail is in itself (individually), the more frequently this detail occurs in proteins. We will see it many times to come.

Finally, let us consider the potential of rotation around the bond between sp^3-hybridized and sp^2-hybridized atoms. Angles of rotation around such bonds are denoted as φ (rotation around N-C^{α}) and ψ (rotation around C^{α}–C′) (Fig. 2.1).

This rotation yields a six-fold (six minimums and six maximums within 360°) potential with rather low barriers (<1 kcal mol^{-1}, see Fig. 2.6) that are

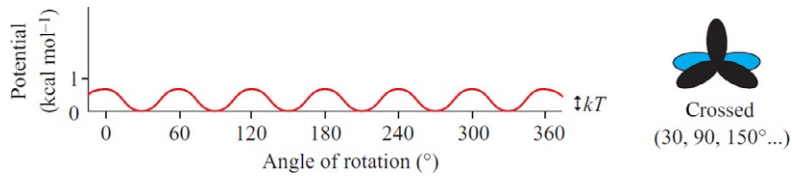


FIG. 2.6 Typical potential of rotation around a single bond between sp^3- and sp^2-hybridized atoms (exemplified by rotation around $H_3C–C_6H_5$). The sp^2-hybridized (light-blue) and the sp3-hybridized (black) electron clouds are shown in the 'crossed' conformation.

of the same order as the energy of thermal fluctuations (which, as we remember, amounts to 0.6 kcal mol^{-1} at room temperature). It is these nearly free rotations around such bonds (N–C^{α} and C^{α}–C′) in the polypeptide main chain that ensure the flexibility of the polypeptide.

Concluding, I have to add that, although potentials of rotation around covalent bonds seem to be rather well established (Halgren, 1995; Levitt et al., 1995; Jorgensen et al., 1996; Wang et al., 2004), some correction of torsional potentials was necessary (Lindorff-Larsen et al., 2010) for the recently achieved progress in reproducing protein folding by molecular dynamics simulations (Shaw et al., 2010).

REFERENCES

Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second ed. Garland Science, New York (Chapter 6).

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman & Co, New York (Part 1, Chapter 2).

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. H. Freeman & Co., New York (Chapters 1, 2, 5).

Halgren, T.A., 1995. Merck molecular force field. I. Basis, form, parameterization and performance of MMFF94. J. Comput. Chem. 17, 490–519.

IUPAC-IUB, 1969. Commission on Biochemical Nomenclature. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969). Biochemistry 9, 3471–3479.

Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118, 11225–11236.

Landau, L.D., Lifshitz, E.M., 1977. Quantum Mechanics (Vol. 3 of A Course of Theoretical Physics). Pergamon Press, Oxford, New York (Section 16).

Levitt, M., Hirshberg, M., Sharon, R., Daggett, V., 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Commun. 91, 215–231.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., Shaw, D.E., 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78, 1950–1958.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co, New York (Chapter 3).

Pauling, L., 1970. General Chemistry. W.H. Freeman & Co, New York (Chapters 3–6).

Schulz, G.E., Schirmer, R.H., 1979 & 2013. Principles of Protein Structure. Springer, New York (Chapter 2).

Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shah, Y., Wriggers, W., 2010. Atom-level characterization of structural dynamics of proteins. Science 330, 341–346.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A., 2004. Development and testing of a general amber force field. J. Comput. Chem. 25, 1157–1174.

This page intentionally left blank

Lecture 3

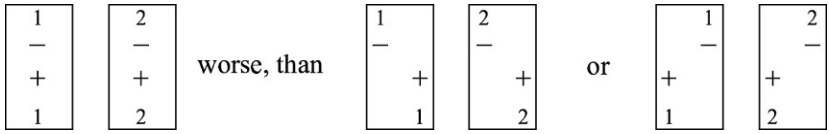
In this lecture, we shall focus on elementary noncovalent interactions between atoms.

First of all, let us recall in what circumstances no covalent bond is formed between approaching atoms. Let me remind you that chemical bonding basically results from delocalization of electrons, that is, their "transition" from one atom to another and back: these electrons are often said to share the same (common) orbital enveloping two or more atoms. However, according to the *Pauli Exclusion Principle* (Landau and Lifshitz, 1977), not more than two electrons can share the same orbital, and they can do this only when their spins (moments of rotation) are oppositely directed (then they are "paired electrons"). Pairing in a common orbital of electrons coming from two different atoms results in a tight covalent bond.

If orbitals of two mutually approaching atoms already bear a pair of electrons each, no covalent bond can emerge. Otherwise, there would be too many electrons in the common orbital—four. But, as stated by the *Pauli Exclusion Principle*, one orbital can bear no more than two electrons. Consequently, a "saturated" orbital with an electron pair on cannot accept extras. Therefore, atoms with saturated electron orbitals repel as they come near enough for their electron clouds to begin to overlap. Such atoms are impenetrable to one another at ordinary (though not at stellar) temperatures.

The same is observed when molecules with no vacant valency approach one another: they repel at a distance between their atoms as short as 2–3 Å.

However, at a greater distance (when electron clouds do not overlap) all atoms and molecules attract each other (London, 1937; Landau and Lifshitz, 1977; Dill and Bromberg, 2010), unless they are charged (we will discuss this later). This attraction is purely quantum in nature. It is connected with coordinated vibrations of electrons in both atoms. The thing is, the coordinated (in the same direction) shift of electrons results in attraction of the atoms (whereas atoms with nonshifted electrons do not attract or repel each other). This becomes clear from the following diagram showing two atoms, 1 and 2, with an electron "−" and nucleus "+" in each.



When considering this diagram, one must bear in mind that electric energy increases with decreasing distance r as $1/r$. The electron shift causes no change

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00003-X

in electron-electron and nucleus-nucleus interactions, but the increase in attraction between electron 1 and nucleus 2, which become close (see the extreme right of the diagram) is greater than the decrease in attraction between electron 2 and nucleus 1. Now recall that electrons are in constant vibration within an atom (they "orbit the nucleus" and cannot fall onto it because of Heisenberg's quantum uncertainty). The above effect provokes a coordination of electron vibrations as the atoms come closer and hence causes attraction of these atoms.

At greater distances, the atomic interactions weaken. As a result, interaction energy decreases as a function of distance r between the centers of the two atoms; it can be shown that the energy decrease is proportional to $(1/r)^6$ (London, 1937; Landau and Lifshitz, 1977).

The total potential of atomic interaction (also called "the energy of van der Waals interaction") is illustrated by Fig. 3.1 and approximately described by the Lennard-Jones potential of the form:

$$U_{\mathrm{LJ}}(r) = E_0\left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6}\right] \tag{3.1}$$

Here r_0 (as can be easily proved by taking the U_{LJ} derivative over distance r) is the distance at which the energy U_{LJ} is at a minimum, and E_0 is the depth of the minimum. The last term that decreases as (const./distance)^6 gives the attraction (minus "−" shows that the corresponding energy decreases with decreasing distance); the term of the 12th power gives the repulsion (it is positive, ie, the corresponding energy increases with decreasing distance).

Eq. (3.1) gives a precise description of the attraction at large distances (when $r \gg r_0$). The repulsion at small distances is described only qualitatively as "very strong and exceeding any attraction when r tends to zero." The approximate character of Eq. (3.1) is demonstrated by the fact that atoms are implied to be spherical (since the described interaction is direction-independent), whereas actually, the atomic electron cloud is not spherical because of projecting p-electrons. In general, only quantum mechanics can provide the correct description of interactions between atoms, but it can strictly calculate only very simple systems like the He atom, H_2^+ ion, or H_2 molecule. All other systems have to be described using approximate "semiempirical" equations such as Eq. (3.1), the form of which is

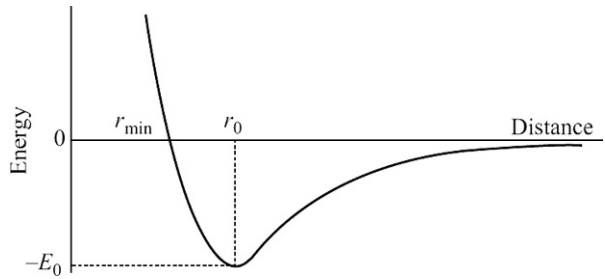


FIG. 3.1 Typical profile of the van der Waals interaction potential.

based on only semiquantitative physical considerations and parameters (in this case, E_0 and r_0) on experiment. Some basic data are given in Table 3.1.

You should notice that Table 3.1 presents not only values of the optimal distance r_0 but also those of the minimum distances r_{min} that exist in the crystalline state. In Fig. 3.1, r_{min} approximately corresponds to the point where energy passes through 0 at a short distance between atoms. Values of r_{min} are helpful in estimating the possibility of a particular chain conformation.

Inner voice: There are a lot of works on deriving potentials of atom-atom interactions (see, eg, Halgren, 1995; Levitt et al., 1995; Jorgensen et al., 1996; Wang et al., 2004). Is not this indicative of the questionable precision of all these potentials?

Lecturer: As to the *form* of the potential (Fig. 3.1), there is no particular disagreement. The estimates of r_{min} are also alike as they are directly measured in crystals. The difference in views concerns values of r_0 and especially E_0. For example, many authors point out that the radius of H in polar N–H and O–H groups is much smaller than that in nonpolar C–H groups. The radii and energies are mainly derived from crystals as well, that is, from their structure (radii) and sublimation heat (energies). However, crystals usually consist of not atoms but molecules, for example, CH_4, C_2H_6… So, when calculating the energy, a question arises as to the contribution of interactions C···C, H···H, and C···H. Different authors answer it in different ways, so sometimes their potentials differ significantly. However, this difference is smoothed out as soon as they come back from atoms to molecules. But it

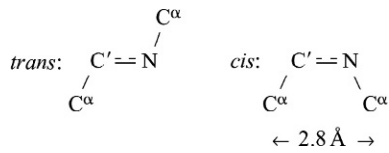
TABLE 3.1 Typical parameters of van der Waals interaction potentials.

Interaction	E_0 (kcal mol^{-1})	r_0 (Å)	r_{min} (Å)	Minimum van der Waals radius of atom (Å)
H···H	0.12	2.4	2.0	H: 1.0
H···C	0.11	2.9	2.4	
C···C	0.12	3.4	3.0	C: 1.5
O···O	0.23	3.0	2.7	O: 1.35
N···N	0.20	3.1	2.7	N: 1.35
CH_2···CH_2	≈0.5	≈4.0	≈3.0	CH_2: ≈1.5

E_0 and r_0 values for interactions between atoms are from Scott and Scheraga (1966); r_{min} values from Ramachandran and Sasisekharan (1968). These values provided the basis for estimating CH_2···CH_2 interaction parameters. The interaction CH_2···CH_2 depends on the relative orientation of these groups; therefore, the tabulated results are approximate. Nevertheless, they are often used to calculate interactions in proteins when H-atoms are "invisible" to X-rays.

is important to remember that when calculating molecular structures, the source of one parameter (say, the energy of the C···C interaction) should not differ from the source of the others (eg, the H···H interaction energy): here, the principle "all-or-none" must be followed to avoid mistakes.

The tabulated values are helpful for understanding why the *trans*-conformation (180 degree) of the C′==N bond is allowed and its *cis*-conformation (0 degree) is disallowed (for all amino acid residues, except Pro, as mentioned earlier): for the C′==N *trans*-conformation, the distance between C^{α} atoms is 3.8 Å, while for its *cis*-conformations (when these atoms are at their closest), it is only 2.8 Å, ie, less than the minimum distance $r_{\min} = 3.0$ Å allowed for the C···C pair.



Inner voice: According to Eq. (3.1), the van der Waals interaction is pairwise. Is it really independent of the environment?
Lecturer: The repulsive part of this interaction is independent because it is due to overlapping of electron clouds of two atoms. The attractive term $(\sim(r_0/r)^6$ in Eq. (3.1)) depends on electrostatic interactions of vibrating electrons; therefore, on average, it decreases (but not very much) with an increase of that component of dielectric permittivity of the media which corresponds to the frequency of these vibrations (Finkelstein, 2007). Actually, the mutual attraction of atoms "1" and "2" can be increased or decreased, depending on the position of atom "3," which is close to them in space (Axilrod and Teller, 1943); the most important in this respect is atom "3" covalently bound to either atom "1" or "2" (Finkelstein, 2007).

Here, it is worth making a comment. Considering the interactions of proteins and other molecules surrounded by water, people often prefer to consider water *implicitly*, as something that has no molecular structure but only changes the value of interaction between the molecules (the use of the medium dielectric constant is the most famous example of this kind). The same is applicable to van der Waals interactions. The approaching of atom "1" to "2" displaces waters (w) from these atoms (and these displaced waters start to interact between themselves). The resulting energy of this approach in water can be presented as $\Delta E_{12} = E_{12} + E_{ww} - (E_{1w} + E_{2w})$, where E_{12}, etc., are approach energies of the corresponding particles in vacuum. The values ΔE for interactions in *implicitly* considered water can be obtained from the crystal dissociation in water in the same way as the in-vacuum energies E are usually obtained from the crystal sublimation (Pereyaslavets and Finkelstein, 2012).

C^{α} atoms of the sequence-neighboring amino acids are rather far apart in space owing to the rigid *trans*-form of the C′==N bond. This provides an opportunity for these neighboring residues to change their conformations almost independently of each other. But inside a residue, rotations over ϕ and ψ angles are interconnected. The "allowed" and "disallowed" conformations of a residue plotted in the (ϕ, ψ) coordinates are called *Ramachandran plots* (Ramachandran and Sasisekharan, 1968) or, to be more exact, Sasisekharan-Ramakrishnan-Ramachandran plots.

Prior to drawing these maps, let us see what conformations are allowed (and what are not) in the case of ϕ (about N–C^{α}) and ψ (about C^{α}–C′) rotations separately.

As we already know, rotation around these bonds (between the sp^3-hybridized C^{α} atom and sp^2-hybridized N or C′) is nearly free. However, in *cis*-conformations (at $\phi = 0$ degree or $\psi = 0$ degree), atoms rotating around these bonds (C'_{i-1} and C'_i for ϕ_i, and N_i and N_{i-1} for ψ_i; see Fig. 3.2; $i-1$, i, $i+1$ are numbers of consecutive residues in the chain) come close to each other, and, because of their repulsion, this conformation may be disallowed, or in other words, sterically prohibited.

$2.9\ \text{Å} < r_{\min}(\text{C}\cdots\text{C}) = 3.0\ \text{Å}$ $\quad$ $2.9\ \text{Å} > r_{\min}(\text{N}\cdots\text{N}) = 2.7\ \text{Å}$

The above scheme shows that minimum distances between the atoms C'_{i-1} and C'$_i$ (ϕ_i angle) and between N_i and N_{i-1} (ψ_i angle) are the same, 2.9 Å.

The 2.9 Å is a bit *less* than $r_{\min} = 3.0$ Å (see Table 3.1) for the C⋯C interaction (so the *cis*-conformation over ϕ_i is disallowed) and a bit *more* than $r_{\min} = 2.7$ Å (but less than the optimal $r_0 = 3.1$ Å) for the N⋯N interaction (so the *cis*-conformation over ψ_i is not prohibited, though strained; however, as it can be shown, even the minor ($\approx \pm 5$ degree) flexibility of the covalent angle N–C^{α}–C′ is sufficient to relieve the strain considerably). If we only had to consider these C′⋯C′ and N⋯N interactions, Ramachandran plots of the prohibited, strained, and allowed conformations would look as shown in Fig. 3.2.

Inner voice: You speak about "prohibited," "strained," and "allowed" conformations. OK, but what does this mean in the energy terms? And how can we understand whether this or that energy effect is significant for a protein or not?
Lecturer: When speaking of significance of energy effects in general, it is useful to remember the following. For an *individual* element (eg, for the above shown turn about a covalent bond or for an amino acid residue),

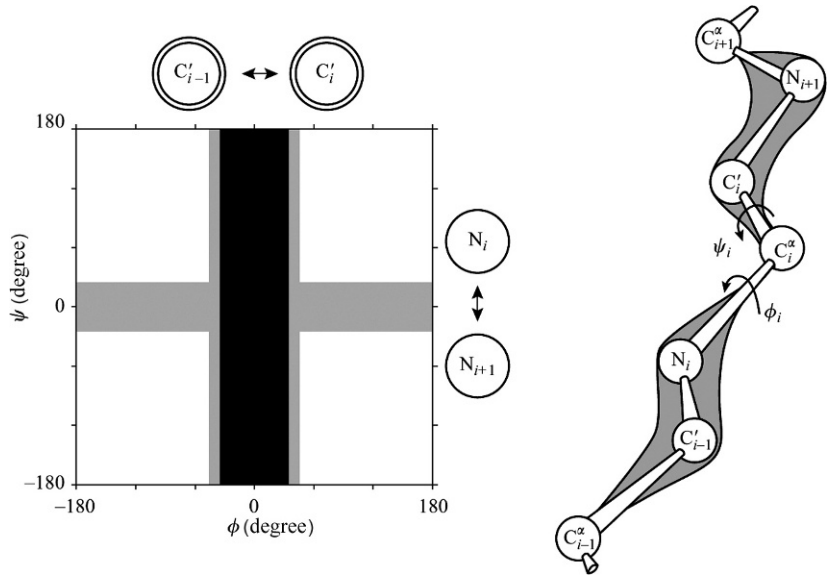


FIG. 3.2 This is how Ramachandran plots of the disallowed ■, somewhat strained ▒, and fully allowed □ (ϕ, ψ) conformations of the fragment $C^\alpha C'N–C^\alpha–C'NC^\alpha$ would look, provided all these atoms had no other atoms attached (right) and atoms of residues $i-1$ and $i+1$ had no interactions between themselves.

the impact of energy below "heat quantum" kT is hardly significant: it is "washed out" by thermal fluctuations. Thus, $kT \approx 0.6$ kcal/mol (at room temperature) is the first threshold worth remembering.

When speaking of significance of the *individual* energy effects for proteins, it is useful also to know that a characteristic structural "reserve of stability" (the difference between free energies of the native and denatured form of a protein) is about 5–10 kcal/mol (as we will see later on). The protein can be "exploded" by an energy "defect" that exceeds 5–10 kcal/mol. Therefore, such "defects" are strongly prohibited, as we will see when discussing statistics of protein structures.

Inner voice: This reminds me of the saying "what's good for General Motors is good for America"… Do you mean that one can say "what's good for a protein's detail is good for the whole protein, and what's bad for a protein's detail is bad for the whole protein"?

Lecturer: Exactly. A stable protein structure must be mostly built from stable elements! The impact of element's energy on statistics of its occurrence in observed protein structures requires a careful analysis which will be done later on. Now, very roughly, I can say the following to conclude this discussion:

"fully allowed" are conformations whose energy exceeds the minimal one by less than one kT;

"strained" conformations are those whose energy exceeds the minimal one by about $kT \approx 0.6$ kcal/mol or a little more;
"prohibited" conformations are those whose energy exceeds the minimal one by several kT (or kcal/mol); and
"strongly prohibited" conformations are those whose energy exceeds the minimal one by many kcal/mol.

Turning back to the Ramachandran plots, we can state that with only C′···C′ and N···N interactions considered, the ϕ, ψ regions for the prohibited, strained, and allowed conformations would look as shown in Fig. 3.2.

Fig. 3.2 shows that, in this case, the ϕ, ψ rotations would be completely independent of each other.

However, the C′ atoms have, in addition, O atoms attached, and the N atoms are, in addition, bonded to H atoms (and in water this H atom is quite rigidly linked by a hydrogen bond to a water molecule, as we will soon see). As a result, the Ramachandran map, that is, plots of disallowed and allowed conformations of the smallest amino acid residue, glycine (with H as its side chain), looks as presented in Fig. 3.3.

Glycine has no massive side chain. All other amino acid residues do have such a chain, and its collision (or, rather, the collision of its C^{β} atom closest to the main

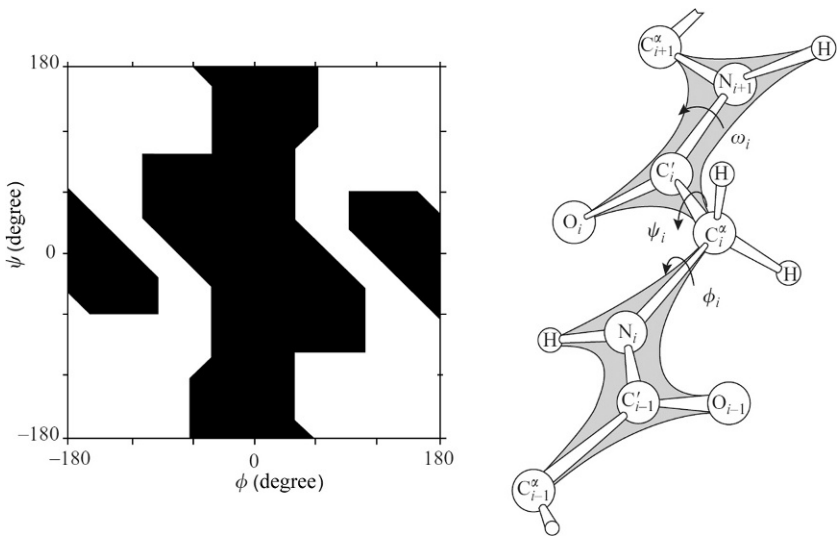


FIG. 3.3 The map of disallowed ■ and allowed □ ϕ, ψ conformations of glycine (Gly) in the protein chain. Angle $\omega = 180$ degree. The contour of allowed regions here, as well as in subsequent figures, is taken from (Finkelstein and Ptitsyn, 1977), where (unlike in Ramachandran and Sasisekharan, 1968) two additional physical factors are taken into account: (i) flexibility of covalent angles (the flexibility of N–C^{α}–C′ is especially important) and that of the angle ω and (ii) hydrogen bonding of NH groups to waters (where the lines N–H···O_{water} are virtually straight). The contours were drawn using r_{min} values listed in Table 3.1.

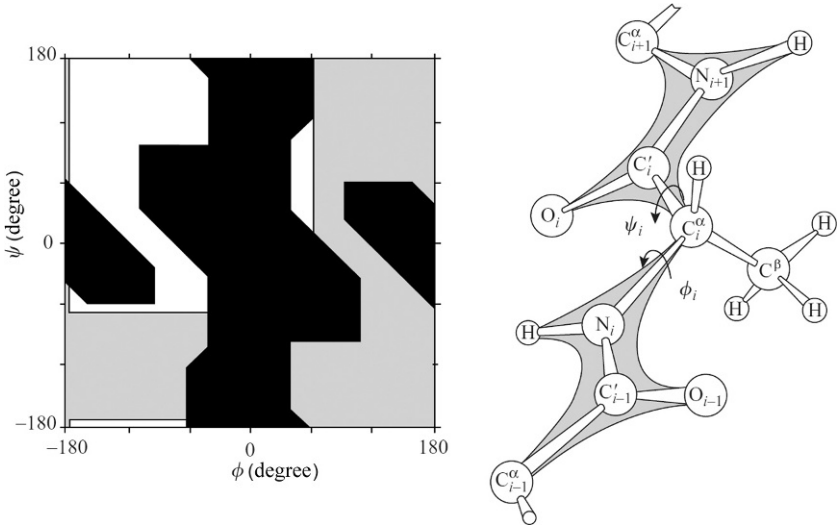


FIG. 3.4 The map of allowed □ ϕ, ψ conformations of alanine (Ala) in the protein chain; ▩, regions allowed for Gly only; ■, regions disallowed by main-chain interactions for all residues.

chain) with the C'_{i-1}-atom accounts for the disallowed ϕ region, while its collision with the N_{i+1} atom accounts for the disallowed ψ region (Fig. 3.4).

The map shown in Fig. 3.4 is for alanine that has a small side chain comprising the $C^{\beta}H_3$ group only. The side chains of all other amino acid residues are larger; they include one or two heavy γ atoms attached to the C^{β} atom. Since these "new" γ atoms (and the still more remote δ, ε, etc.) are far from the main chain, their effect on the Ramachandran map is only minor.

More precisely, in a small region (left white in Fig. 3.5), γ atoms have no collisions at all with the main chain, whereas in other conformations allowed for alanine there are such collisions for side chains with larger (C or S, but also O) γ atoms. These collisions between γ atoms and the main chain are most significant for valine, isoleucine, as well as threonine, which have two large γ atoms each. Therefore, these three residues must be somewhat strained (Leach et al., 1966) outside the white circle shown in Fig. 3.5.

Next, let us consider the Ramachandran plot for the imino acid proline, where the ϕ angle is nearly fixed at -70 degree with a ring built up by the Pro side group linked to the N-atom of its main chain, while rotation over ψ is similar to that of alanine. As a result, the allowed conformations of proline are accommodated by the white band in the Ala's Ramachandran plot, Fig. 3.6.

The N atom-bound Pro ring also diminishes the region of allowed conformations of the residue preceding proline in the polypeptide chain (Fig. 3.7).

Finally, let us see whether all these theoretical considerations agree with the observed (by X-ray crystallography and NMR spectroscopy) conformations of amino acid residues in proteins. In Fig. 3.8, these observed conformations are

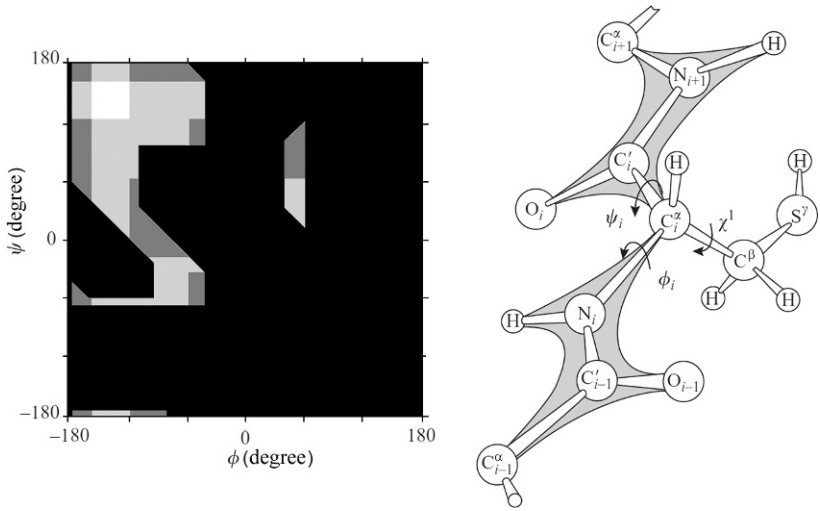


FIG. 3.5 The map of disallowed ■ and allowed (□, ▒, ▓) ϕ, ψ conformations of larger residues in the protein chain. □, the region where all three χ^1-rotamers of the side chain C^{γ} (or S^{γ}) atom are allowed; ▒, the region where two χ^1-rotamers of the C^{γ} atom are allowed and one is disallowed; ▓, the region where only one χ^1-rotamer of the C^{β} atom is allowed and two others are disallowed. *(Adapted from Finkelstein, A.V., Ptitsyn, O.B., 1977. Theory of protein molecule self-organization. I. Thermodynamic parameters of local secondary structures in the unfolded protein chains.* Biopolymers *16, 469–495.)*

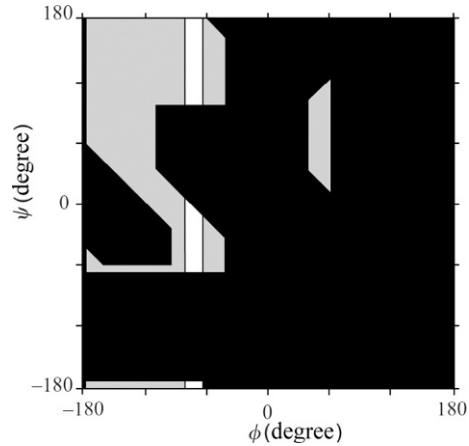


FIG. 3.6 The map of allowed □ Pro conformations plotted against allowed Ala conformations ▒; ■, the region of disallowed conformations for both residues.

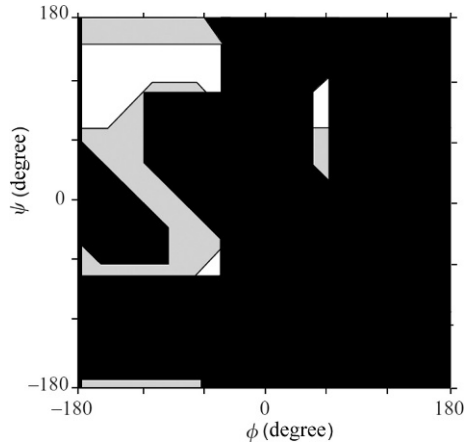


FIG. 3.7 The region of allowed □ conformations of an Ala residue that precedes Pro in the protein chain. If not the following Pro, ■ would also be included in the region of allowed Ala's conformations.

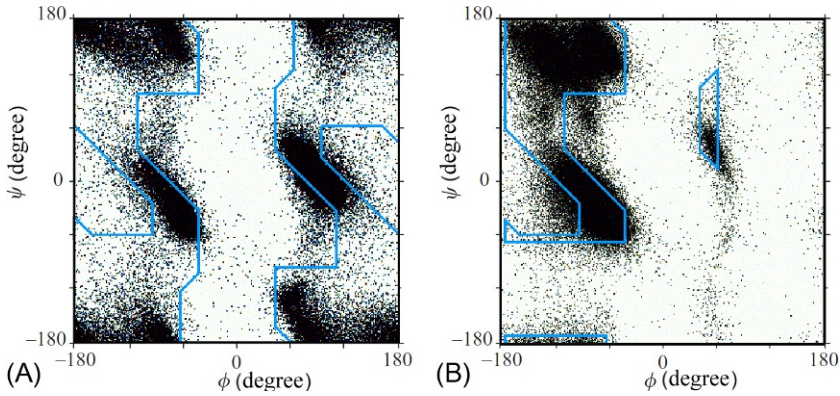


FIG. 3.8 Observed conformations (dots) of glycine (A) and of other amino acid residues (B) in nonhomological proteins taken from the Protein Data Bank (Berman et al., 2012). The sterically allowed regions are contoured.

plotted against the contours of regions theoretically allowed for glycine, on the one hand, and for alanine and all other residues, on the other hand. As seen, the agreement is quite good. We see that the "sterically allowed" regions accommodate the majority of the experimental points. Some parts of these regions are more populated than others. Later on, we will see that these parts are related to secondary structures. We also see that some points are in the "sterically prohibited" regions. This is not surprising, since the regions that we call "sterically prohibited" are those of high energy—not infinitely high, for sure, as that would

cause complete prohibition—but higher than the minimum conformational energy by, say, a couple of kilocalories per mole. In other words, a protein has to spend some energy to drive its residue into such a region. We see that it is able to do so, although it rarely does.

In the course of these lectures, we will learn the general rule: strained, high-energy elements are rare, though sometimes observed. This is not surprising since a stable protein must contain—mostly, but not exclusively—stable components.

REFERENCES

Axilrod, B.M., Teller, E., 1943. Interaction of the van der Waals' type between three atoms. J. Chem. Phys. 11, 299–300.

Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L., 2012. The Protein Data Bank at 40: Reflecting on the past to prepare for the future. Structure 20, 391–396. http://www.wwpdb.org/.

Dill, K.A., Bromberg, S., 2010. Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience, second ed. Garland Science, New York (Chapter 24).

Finkelstein, A.V., 2007. Average and extreme multi-atom van der Waals interactions: Strong coupling of multi-atom van der Waals interactions with covalent bonding. Chem. Central J. 1, 21.

Finkelstein, A.V., Ptitsyn, O.B., 1977. Theory of protein molecule self-organization. I. Thermodynamic parameters of local secondary structures in the unfolded protein chains. Biopolymers 16, 469–495.

Halgren, T.A., 1995. Merck molecular force field. I. Basis, form, parameterization and performance of MMFF94. J. Comput. Chem. 17, 490–519.

Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118, 11225–11236.

Landau, L.D., Lifshitz, E.M., 1977. Quantum Mechanics. A Course of Theoretical Physics, vol. 3 Pergamon Press, Oxford, New York (Sections 62, 89).

Leach, S.J., Némethy, G., Scheraga, H.A., 1966. Computation of the sterically allowed conformations of peptides. Biopolymers 4, 369–407.

Levitt, M., Hirshberg, M., Sharon, R., Daggett, V., 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Commun. 91, 215–231.

London, F., 1937. The general theory of molecular forces. Trans. Faraday Soc. 33, 8–26.

Pereyaslavets, L.B., Finkelstein, A.V., 2012. Development and testing of PFFsol_1, a new polarizable atomic force field for calculation of molecular interactions in implicit water environment. J. Phys. Chem B 116, 4646–4654.

Ramachandran, G.N., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 283–438.

Scott, R.A., Scheraga, H.A., 1966. Conformational analysis of macromolecules. III. Helical structures of poly-glycine and poly-L-alanine. J. Chem. Phys. 45, 2091–2101.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D., 2004. Development and testing of a general amber force field. J. Comput. Chem. 25, 1157–1174.

This page intentionally left blank

Lecture 4

So far we have not taken the aqueous environment of proteins into account. It's high time we bridged the gap.

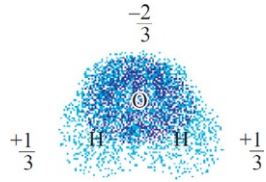
Water is a peculiar solvent. First, it boils and freezes at abnormally high temperatures compared with those typical for substances of similarly low molecular weight. Indeed, water (H_2O) boils at 373 K and freezes at 273 K, while O_2 boils at 90 K and freezes at 54 K; H_2 boils at as low a temperature as 20 K and freezes at 4 K; CH_4 boils at 114 K, etc. The fact that ice and water structures are heat-resistant suggests some strong bonding among water molecules.

The bond responsible for this effect is specifically that between O- and H-atoms of H_2O. This bond is called a *hydrogen bond*.

Hydrogen bonds are not only observed in water. They invariably occur when a hydrogen atom approaches some electronegative (ie, electron attracting) atom while being chemically bonded to another electronegative atom, as exemplified by the O–H:::O, N–H:::N bonds. But, for instance, a C–H group is not involved in perceptible hydrogen bonding since the electronegativity of the C atom is insufficient.

The solvent properties of water are dominated by strong hydrogen bonds (Sokolov, 1955; Pauling, 1970). The hydrogen bonding between water molecules is nearly completely electrostatic in nature. It is connected with electrons and charges, but not with the nuclei of the hydrogen atoms (unlike F–H:::F bonds involving atoms F that are much more electronegative than O (see Sokolov, 1955); as is shown by the close similarity of boiling and melting parameters of light (H_2O) and heavy (D_2O) waters in spite of a twofold difference in mass between D and H nuclei.

The water molecule is polar. This implies small ("partial") charges on its atoms: negative on O and positive on H. The distribution of charges and electron clouds at these atoms appears as follows:



Here, the density of dots reflects the density of the electron cloud, and numerals indicate the partial charges on the atoms. These are expressed in fractions of the proton charge which, naturally, in these units amounts to +1 while the electron charge is −1.

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00004-1

An electronegative oxygen draws the electron clouds off the neighboring hydrogens, thereby causing polarization of atoms. As a result, H atoms acquire partial positive charges, while there is a negative charge on O in the water molecule.

As you will remember, in a vacuum, the energy of interaction of charges q_1 and q_2 at distance r is

$$U = \frac{q_1 q_2}{r} \tag{4.1}$$

In a vacuum, interactions between charges are very strong. The energy of interaction of two single charges (ie, proton or electron charges) at a distance of 1 Å is nearly 330 kcal mol^{-1} (keep this figure in mind: we will use it for different estimates later), while at a more realistic distance of 3 Å (with van der Waals repulsion of atoms taken into account) this energy is about 110 kcal mol^{-1}. The energy of single-charge interaction is the typical energy of a chemical bond; it is hundreds of times higher than the typical thermal energy kT or the typical energy of van der Waals interactions between atoms.

The partial charges of water molecules are still lower than unity, and therefore, their interaction is weaker: at a distance of 3 Å its energy is about 10 kcal mol^{-1}; however, this energy is sufficient to distort the electron envelopes of H-atoms by H to O attraction. Hydrogen atoms are most sensitive in this respect: their single-electron envelope is drawn towards O and therefore undergoes the distortion most easily. It takes much more energy to distort, say, the electron envelope of oxygen that has eight "own" electrons and a share of the single electron of both hydrogens of the water molecule.

It is the ease of distortion of the electron cloud of a hydrogen atom that turns a normal electrostatic interaction into a hydrogen bond. This is true for all hydrogen bonds, among which those of interest to us are: O–H:::O, N–H:::O, N–H:::N.

Thus, a hydrogen atom has the thinnest cloud, whose significant distortion results from attraction between the partial positive charge of hydrogen and the partial negative charge of oxygen (or nitrogen). This gives distances between H and O (or N) nuclei as small as 1.8–2.1 Å (reported for crystals of small molecules) instead of the 2.35–2.75 Å typical for van der Waals interactions discussed in the previous lecture. Therefore, it is accepted that the van der Waals radius of H in the O–H or N–H group is about 30% less than that of H in the C–H group (Levitt et al., 1995).

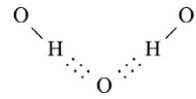
←1.8 Å→	←1.9Å→	←2.1Å→
O—H : : : O	N—H : : : O	N—H : : : N
← 2.8 Å →	← 2.9 Å →	← 3.1 Å →

This close approach yields a *hydrogen bond* (or *H-bond*). The H atom (to be more exact, the O–H or N–H group) is called the *donor* of the hydrogen bond,

and the O or N atom towards which the hydrogen moves, is called the *acceptor* of the hydrogen bond.

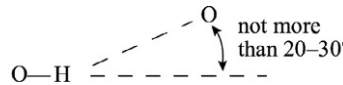
Note that in crystals, H-bonding occurs when the distance between O- and/or N-atoms of the donor and acceptor is about 3 Å (eg, in ice it is 2.8 Å). This is similar to the optimal van der Waals distance between O and/or N atoms, ie, the presence of the mediating H atom does not increase the distance between these atoms of the donor and acceptor, as it pushes them, not apart, but together.

Each H-bond has one donor and one acceptor. The hydrogen atom almost always acts as a donor of only one H-bond, while the oxygen atom may participate as an acceptor in two H-bonds simultaneously, thereby causing "fork-like" H-bonding:



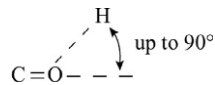
Since the "fork-like" H-bonding implies a short distance between two H atoms (with +1/3 charge on each), its total energy is less than double the energy of a single H-bond.

Unlike van der Waals interactions, H-bonding is rather orientation-sensitive, especially as concerns the orientation of the donor group (Ramakrishnan and Prasad, 1971; Finkel'shtein, 1976). Usually, a valence bond of the donor is directed at the acceptor atom (O or N) to be involved in the hydrogen bond:



It seems that this effect is due to repulsion between O atoms and a very small radius of H atom, when H is in the O–H (or N–H) group.

Orientation of the acceptor group is of considerably less importance for the H-bond:



The H-bond energy is about 5 kcal mol^{-1}. This estimate results from comparison of experimental evaporation heats of similar compounds, some of which are capable of H-bonding, while others are not. For example, the evaporation heat of dimethyl ether, $H_3C–O–CH_3$, is about 5 kcal mol^{-1}, and that of ethanol, $CH_3–CH_2–OH$, is about 10 kcal mol^{-1}. These compounds consist of the same

atoms (ie, their van der Waals interactions are nearly the same), but ethanol is capable of H-bonding, while dimethyl ether is not (as it lacks the O–H group). Each O–H group can participate as donor in only one H-bond, and its O atom can accept this bond. Since each H-bond is supposed to have one donor and one acceptor, there is only one H-bond per ethanol molecule, that is, an H-bond "costs" about 10 kcal mol^{-1} (*ethanol*) − 5 kcal mol^{-1} (*ether*) = 5 kcal mol^{-1}.

The same estimate follows from the value of ice evaporation heat (680 cal g^{-1} = 680 cal (1/18 mol)$^{-1}$ = 12 kcal mol^{-1}). Here, a couple of kilocalories per mole are the share of van der Waals interactions, as seen from the evaporation heat of small molecules such methane (CH_4) or O_2. The remaining 10 kcal mol^{-1} are for H-bonding. In ice, there are two H-bonds (Fig. 4.1) per molecule of H_2O, since its two O–H groups can serve as donors for two H-bonds (and as many can be accepted by its oxygen). Again, the "cost" of one H-bond appears to be about (10 kcal mol^{-1})/2 = 5 kcal mol^{-1}.

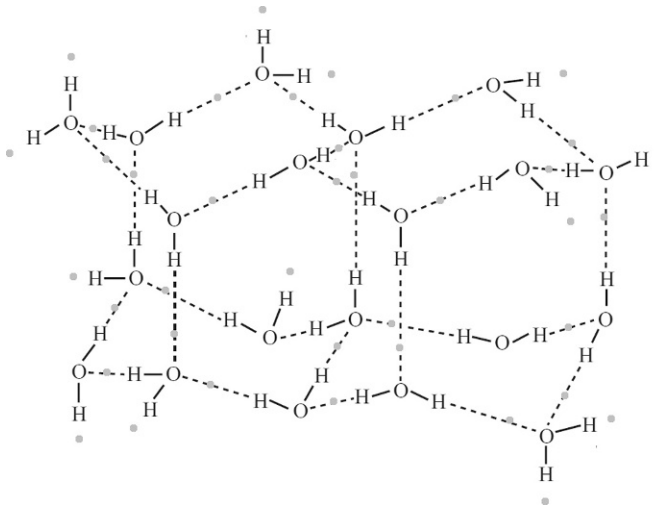


FIG. 4.1 Normal ice ("ice I_h"; there are also other forms of ice, but they can be stable at very high pressure only). The *continuous lines* show covalent bonding, the *dashed lines* show H-bonding. As seen, the openwork structure of ice has small cavities surrounded by H_2O molecules. The drawing is adapted from Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W. H. Freeman & Co., New York (Chapter 4), with minor modifications. In this drawing, each H atom is unambiguously bound to one O atom. However, as shown by X-ray analysis (Madura, 1994), each H atom can occupy *two* positions with an equal probability of 50%: *either* close to an O atom with which it is covalently bound (as shown in the picture), *or* close to an O atom with which it forms an H-bond (the latter, alternative position of H atom, is shown by a small gray spot in the picture). Transition of an H atom from one O atom to another is relatively easy. In ice, this transition is connected with rearrangement of all covalent and H-bonds of these two O atoms and of all their neighbors. As a result, the entire network of covalent and H-bonds in the ice can fluctuate greatly, which leads to its abnormally high polarizability: its permittivity is 97 at 0°C, exceeding that of liquid water (88 at 0°C) (Lide, 2005).

The structure of normal ice is determined by H-bonds (Fig. 4.1): it is good for their geometry (O–H is directed at O), although not so good for close van der Waals contacts between water molecules. In ice, water molecules envelop tiny (smaller than H_2O molecules) pores, thereby giving it an openwork structure. This results in two well-known phenomena: (1) ice is not as dense as water, it floats, and (2) under strong pressure (eg, caused by skate blades), ice melts. It is also the case that the abnormally high permittivity of ice is indicative of easy rearrangement of the H-bonds (and covalent bonds!) in it; this is illustrated by Fig. 4.1.

Inner voice: I can readily accept easy rearrangement of H-bonds. But "easy rearrangement of covalent bonds" sounds strange indeed … Normally, covalent bonds are stable in the absence of a catalyst!
Lecturer: Easy rearrangement of bonds formed by hydrogens with electronegative atoms must not surprise you, because, from chemistry, you must be already familiar with easy transfer of H atoms from acids to water in the absence of any catalyst (except environment which consists of electronegative atoms and hydrogens bonded to them) … Such an easy rearrangement of the covalent bonds relates also to coordinate bonds of electronegative atoms with certain metal ions, which we will consider soon.

The majority of H-bonds existing in ice (Fig. 4.1) persist in liquid water. This follows from the low melting heat of ice (80 cal g^{-1}) as compared with water boiling heat (600 cal g^{-1} at 0°C and 540 cal g^{-1} at 100°C). We might think that only as many as $80/(600+80) = 12\%$ of all H-bonds existing in ice, break down in liquid water. However, this picture—with some H-bonds broken while others persist—is not quite true: rather, in liquid water all H-bonds exist but become slightly loose.

This is well illustrated by the experimental results shown in Fig. 4.2.

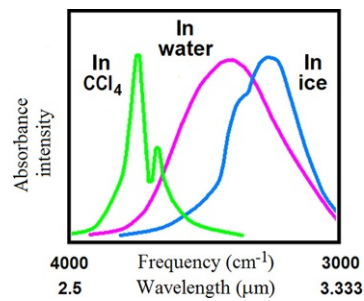


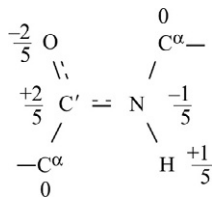
FIG. 4.2 Typical IR (infrared) absorption spectra for O–H groups in ice, CCl_4 solution and in liquid water. *(Adapted from Greinacher, E., Lüttke, W., Mecke, R., 1955. Infrarotspektroskopische Untersuchungen an Wasser, gelöst in organischen Lösungsmitteln. Z. Electrochem. 59, 23–31; Zolotarev, V.M., 1970. Optical constants of ice I in a broad infrared spectrum. Opt. Spectrosc. (in Russian). 29, 1125–1128.)*

Here, Curve "in ice" shows the maximum of the infrared (IR) absorption spectrum for O–H groups in ice (where all H-bonds are saturated); Curve "in CCl_4" illustrates this maximum for the O–H-groups of separate water molecules dissolved in CCl_4 (where no H-bonding occurs because of the extreme dilution); and Curve "in water" shows the absorption spectrum for liquid water. Suppose liquid water contained two types of O–H groups: those participating and those not participating in H-bonds. Then the former would vibrate with frequencies typical for ice (where they have been involved into H-bonding), while the latter would vibrate like those in water molecules dissolved in CCl_4 (where no H-bonding has occurred). Then the IR absorption spectrum for liquid water would be double-peaked to reflect the two types of O–H groups and their two typical frequencies, since the vibrational frequency of a group is equal to its light absorption frequency. However, no such "double-peaked" picture is actually observed. Instead, Curve "in water" presents one broad peak that goes from the peak on Curve "in ice" to that of Curve "in CCl_4." This means that in liquid water, all O–H groups are involved in H-bonding and all resulting H-bonds are loose but in different ways.

Strictly speaking then, the model with some H-bonds persisting in liquid water and others broken is incorrect. However, people often use it, owing to its simplicity and convenience in describing the thermodynamic properties of water—and we may use it as well, although its drawbacks must be kept in mind.

Now let us concentrate on interactions between the protein chain and water molecules.

As with water molecules, the backbone of a protein chain is polar. To be more exact, it is its peptide groups that are polar. The net charge of the protein chain backbone is 0, and the distribution of charges on its atoms (again, the charge on each atom is expressed in fractions of the proton charge) is as follows:

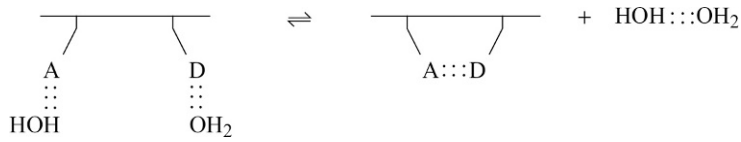


Partial charges are acquired by some side-groups too, eg, by that of Ser (its side-group, -CH_2-OH, is similar to ethanol). Charged amino acid residues are even more polarized: a charge of −1 is typical for the acidic side-groups of Asp and Glu when they are ionized (at about neutral pH), while +1 is characteristic of the ionized basic side-groups of Arg, Lys, and His.

Both main-chain peptide groups and polar side-groups participate as donors and acceptors in hydrogen bonds. They can be—and mostly are—involved in H-bonds formed among themselves or to water molecules: since the H-bond

energy (5 kcal mol^{-1}) is about an order of magnitude higher than that of thermal movement, H-bonds are mostly preserved by these movements.

When an H-bond between donor (D) and acceptor (A) is formed within a protein molecule in the aqueous environment, it replaces two hydrogen bonds between the protein and water molecules (that existed earlier, since an H-bond is too expensive to be neglected), and an additional bond between the two freed water molecules is formed (for the same reason):



The energy balance of this reaction is close to 0: two bonds yield two bonds. However, the *entropy* of the water molecules increases since they are no longer bonded to the protein chain but only mutually H-bonded and free to go anywhere (and the entropy results just from the freedom of movement). This entropy increase caused by the water dimer release is approximately equal to an entropy increase resulting from the molecule H_2O transition from ice to liquid water (in both cases one particle becomes free in its movements).

The entropy difference resulting from one molecule H_2O transition from ice to liquid water can be estimated as follows. At the melting-point (for ice, at 0°C, ie, 273 K), the increase in entropy caused by melting is known to fully compensate for the corresponding decrease in energy. And we know the value of this decrease—it is $80 \text{ cal g}^{-1} \times 18 \text{ g mol}^{-1} = 1440 \text{ cal mol}^{-1}$ (18 g mol^{-1} is the molecular weight of H_2O).

Thus, owing to this increase of water entropy, the free energy of the "protein in water" system decreases by about 1.5 kcal mol^{-1} of emerging intra-protein hydrogen bonds D:::A.

This free energy released by waters may fully or partially compensate the free energy increase that results from decreasing conformational entropy of the chain during the D:::A bonding. As we will see later, a decrease in free energy of the water molecules nearly compensates for the entropy of fixation of the amino acid residue conformation required for H-bonding (N–H:::O) in secondary structures of the polypeptide chains (as a result, in water, the regular secondary structure of polypeptides is just at the edge of stability).

These awful words "entropy" and "free energy" have been articulated … My experience shows that a biologist has normally knows about entropy (that it is a measure of the multitude of possible states, a measure of disorder) but is uncertain about what free energy is … As we often appeal to the concept of free energy (which is a measure of stability), I would like to devote a few minutes to it just now, and later, when the need arises, to discuss it (and entropy too) in more detail.

Let us start with a simple example: suppose a molecule can have two states, "a" and "b"; "a," when it is here, in this room, 200 m above sea level; "b," when it is in the Dalai Lama's monastery, in The Himalayas, 5 km above sea level. What is the relationship of the probabilities of these two states provided that (1) temperature T is the same in both places, and (2) we watch the molecule long enough for it to visit both places?

As stated by the well-remembered Boltzmann formula:

$$\text{Probability of being in the state of energy } E \propto \exp\left(\frac{-E}{kT}\right) \tag{4.2}$$

Physically, the sense of this formula is that the heat of the medium (ie, collisions with other molecules) excites our molecule to a certain extent (proportional to the medium temperature T, on average) and thereby enables it to enter the region of more or less high energies. (All this will be discussed later in more detail but at the moment, I am taking the liberty of believing that you remember this formula.) Still let me remind you that k is Boltzmann's constant, and T is the absolute temperature in Kelvin (K) (counted from the "absolute zero," so OK $= -273.15$°C, and "$\propto$" means "is proportional to").

However, it is better to deduce the Boltzmann formula and not rely on your memory! We will deduce it for at least the case that is of interest to us here, that is, for distribution of gas molecules over height, when their energy is described as $E(h) = mgh$, where m is mass of a gas molecule, g is acceleration of gravity, and h is height.

As stated by Clapeyron-Mendeleev law, the pressure of an ideal gas $P = nkT$, where n is the number of gas molecules per unit volume. If T remains unchanged at any h, then $\mathrm{d}P/\mathrm{d}h = (\mathrm{d}n/\mathrm{d}h)kT$. On the other hand, when considering a gas column of unit cross-section, we see that $\mathrm{d}P = (mgn)(-\mathrm{d}h)$, since the weight of the gas pressing down on the unit cross-section decreases by $(mgn)\mathrm{d}h$ when the height grows by $\mathrm{d}h$. Therefore, $\mathrm{d}P/\mathrm{d}h = (\mathrm{d}n/\mathrm{d}h)kT = -mgn$. Hence, $\mathrm{d}n/\mathrm{d}h = -(mg/kT)n$, or $\mathrm{d}[\ln(n)]/\mathrm{d}h = -mg/kT$, that is, $n \propto \exp(-mgh/kT) = \exp(-E(h)/kT)$.

Thus, as applied to the problem in question (What is the relationship of probabilities for a molecule to be at different heights?), the Boltzmann formula reduces to the barometrical relationship

$$\begin{aligned}&[\text{Probability of being at a height "b"}] \text{ relates to } [\text{Probability of}\\ &\text{being at a height "a"}] - \text{as } \exp\left(\frac{-E_\mathrm{b}}{kT}\right) \text{relates to } \exp\left(\frac{-E_\mathrm{a}}{kT}\right)\end{aligned} \tag{4.3}$$

where E_a, energy of the molecule in the state "a" (ie, "here"); E_b, in the state "b" ("at a height of 5 km"); T, absolute temperature (for simplicity, as mentioned above, is assumed to be invariable with height).

Because of gravity, "here" the energy of the molecule is lower than "at a height of 5 km"; so, according to Boltzmann, the molecule will stay "at a height of 5 km" for a shorter time than "here" (the time will be 1.5–2 times shorter).

Make the calculations yourselves by taking $T = 300$ K and recalling that $E = mgh$, where m = average mass of an air molecule (≈ 30 Da $= 30$ g mol^{-1}), $g \approx 10$ m s^{-2} (gravitational acceleration), $h \approx 5$ km (height difference) and Boltzmann's constant $k = 1.38 \times 10^{-23}$ J degree^{-1} particle^{-1} $= 0.33 \times 10^{-23}$ cal degree^{-1} particle^{-1} (since $J =$ kg m^2 s^{-2} $= 0.24$ cal), ie, ≈ 2 cal degree^{-1} mol^{-1} (since 1 mol $= 6 \times 10^{23}$ particles). As you may remember, $R = 2$ cal K^{-1} mol^{-1} is the "gas constant."

In other words, at a height of 5 km the molecules will be about two times less numerous than here. Or rather, this will be the case *for equal volumes*, for example, your lungs (as you may easily ascertain by breathing at different heights). However, *in total*, the molecules in the Dalai Lama's monastery are much more numerous than they are here just because the monastery is much larger than our room. That is, over there, the molecule may have more positions, since for a freely flying molecule the number of positions is proportional to the room volume. In this case, physicists would say that in the monastery the number of *microstates* of the molecule is far greater than in our room. So, the probability that our molecule is *somewhere* in the Dalai Lama's monastery relates to the probability that it is *somewhere* in this room as

$$\begin{aligned}&[\text{Probability of being somewhere in volume "b"}]\\ &:[\text{Probability of being somewhere in volume "a"}] = \left[V_\text{b}\ \exp\left(\frac{-E_\text{b}}{kT}\right)\right]\\ &:\left[V_\text{a}\ \exp\left(\frac{-E_\text{a}}{kT}\right)\right]\end{aligned} \tag{4.4}$$

where V_a is the volume of "a" ("our room") and V_b is the volume of "b" ("his monastery"). From elementary college maths you will remember that V can be presented as $\exp(\ln V)$, so the above formula can be written as

$$\begin{aligned}&[\text{Probability of being somewhere in volume "b"}]\\ &:[\text{Probability of being somewhere in volume "a"}]\\ &= \left[\exp\left(\frac{-E_\text{b}}{kT} + \ln\ V_\text{b}\right)\right] : \left[\exp\left(\frac{-E_\text{a}}{kT} + \ln\ V_\text{a}\right)\right]\\ &= \left[\exp\left(\frac{-(E_\text{b} - T \times k\ \ln\ V_\text{b})}{kT}\right)\right] : \left[\exp\left(\frac{-(E_\text{a} - T \times k\ \ln\ V_\text{a})}{kT}\right)\right]\end{aligned} \tag{4.5}$$

The last expression looks very much like Eq. (4.3), the Boltzmann equation, but it is applicable not to a volume *unit* but to the *total* volume of a system, and—note carefully—it has $E - T \times k \ln V$ instead of E.

It is the value $F = E - T \times k \ln V$ that is called *free energy* of our molecule of air in a given volume V at temperature T. And the value $S = k \ln V$ is called the *entropy* of our molecule in the volume V (which, in our case, is proportional to the "number of accessible states" of our molecule).

In the general case, entropy S is simply equal to $k \times$ [logarithm of the number of accessible states]. And free energy F relates to energy E, entropy S, and temperature T according to the general equation

$$F = E - TS \tag{4.6}$$

Of two states, the more stable (ie, more *probable*) is the one having a lower free energy:

$$\begin{aligned}&[\text{Probability of being somewhere in volume "b"}]\\ &:[\text{Probability of being somewhere in volume "a"}]\\ &= \exp\left(\frac{-F_b}{kT}\right) : \exp\left(\frac{-F_a}{kT}\right) = \exp\left[-\frac{(F_b - F_a)}{kT}\right]\end{aligned} \tag{4.7}$$

In other words, a more probable, that is, a *more* stable, state of the system is that with a lower F, and the *most* stable state of a system (at a given temperature and volume of this system) corresponds to *the minimum free energy F*.

Thus, the "free energy" is a natural generalization of the regular "energy" for the case when *the system exchanges heat* with the environment. Let me remind you that if a body is *not* excited by environmental heat, its stable state corresponds to its minimum energy (or simply, everything that can fall down—eventually falls down). When excited by environmental heat, molecules of the system start acquiring numerous states of a higher energy (ie, the entropy of this system, of its movements, increases), and as a result, air molecules fly and do not drop onto the ground.

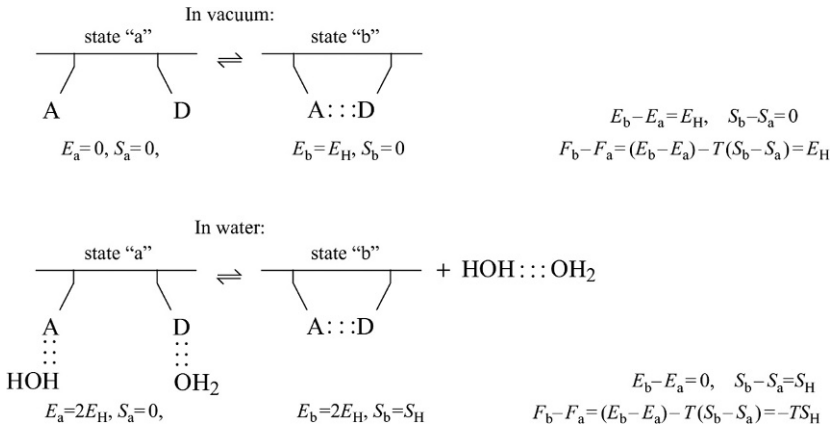
This can also be put as follows:

A change in the energy, $E_b - E_a$, is the work required to transfer a body from state "a" into state "b" when there is no heat exchange with the environment. And a change in the free energy $F_b - F_a$ is the work required to transfer a body from state "a" into state "b" *when the body keeps exchanging heat* with the environment.

Let us now descend from The Himalayas to proteins. So, what is the balance of energy, entropy and free energy in the previous example of H-bonding in the protein chain?

To visualize this, let us compare how the process goes under different conditions (see the scheme below). There, $E_H < 0$ is the H-bond energy, and $S_H > 0$ is the entropy of movements and rotation of a free body, ie, of the free dimer HOH:::OH_2. The H-bonds—between water molecules or between water and protein molecules—are stable when $E_H < -TS_H < 0$ (and if $E_H > -TS_H$ the H-bonding is unstable, and what we deal with is not liquid water but water vapor).

A comparison of the diagrams given earlier shows that H-bonds within the protein chain surrounded by water display a lower stability than when in a vacuum. Indeed, in water, the H-bond free energy is $F_b - F_a = -TS_H$, ie, it is smaller in absolute value than in a vacuum, where $F_b - F_a = E_H$.



I would like to emphasize again that the reason for this decrease in the H-bond stability is that *in water* an H-bonding within the protein chain *replaces* the H-bonding between the chain and water. For the same reason, the hydrogen bonds that stabilize protein structure *in water* are entropic but not energetic in nature: the energies of the two states of the chain (with and without the intra-chain H-bond) are approximately equal, and of these two, the state with a higher entropy (with a greater number of microstates) is more stable. And a free water molecule has a greater number of microstates (ie, a greater number of positions in space) than a bound molecule.

I call your attention to the following: in the protein chain (surrounded by water molecules) hydrogen bonds are entropic and *not* energetic in nature just *because* the energy of H-bonding is extremely high! Consequently, donors and acceptors that are "free" of bonding *within the protein* are not really free of *any* bonding, as they participate in H-bonds to water molecules. The water molecules released from the protein during H-bonding inside the chain immediately bind to one another, thereby compensating for the energy, such that the free energy gain of intra-protein H-bonds occurs only because of the increasing number of possible microstates of the released water molecules. While it is true that to bond to one another, the water molecules have to sacrifice a part of their gained freedom (entropy), it is better to lose a small entropy than a large energy.

Two facts determine the behavior of water as a specific solvent: (1) water molecules are strongly H-bonded to one another; (2) this H-bonding occurs only at a certain mutual orientation of water molecules. A variety of interesting effects are thereby caused. These will be the subjects of Lectures 5 and 6.

REFERENCES

Finkel'shtein, A.V., 1976. Stereochemical analysis of the polypeptide chains secondary structure by Courtauld space-filling models. II. Hydrogen and hydrophobic bonds. Mol. Biol. (Mosk) 10, 724–730.

Levitt, M., Hirshberg, M., Sharon, R., Daggett, V., 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Commun. 91, 215–231.

Lide, D.R., 2005. Section 6 "Fluid properties: permittivity". Section 12 "Properties of solids: permittivity". In: CRC Handbook of Chemistry and Physics on CD. CRC Press, Boca Raton, FL.

Madura, J.D., 1994. Ice coordinates. http://www.ibib_lio.org/water/wsn_archive/msg00037.html.

Pauling, L., 1970. General Chemistry. W.H. Freeman & Co, New York (Chapter 12).

Ramakrishnan, C., Prasad, N., 1971. Study of hydrogen bonds in amino acids and peptides. Int. J. Pept. Protein Res. 3, 209–231.

Sokolov, N.D., 1955. Hydrogen bond. Phys. Usp. (Adv. Phys. Sci., Moscow, in Russian) 57, 205–278.

Lecture 5

Hydrophobicity is a phenomenon that occurs only in an aqueous environment. It plays a very important role in formation and maintenance of protein structures (Bresler and Talmud, 1944; Kauzmann, 1959; Tanford, 1980; Cantor and Schimmel, 1980; Nelson and Cox, 2012).

However, before considering the hydrophobicity, in the initial part of this lecture, I would like to talk on *thermodynamics*. This will be useful for further consideration of water as a specific solvent. The focus of this consideration will be the free energy of immersing various molecules in water.

To study the free energy of immersion of a molecule in water, we take a closed tube, half of which is filled with water and the other half with vapor, and watch how introduced molecules are distributed between these phases.

As we have learned, the difference between free energies F defines a more favorable state of the system (in this case, a more favorable position of the studied molecule) according to the formula:

$$\frac{[\text{Probability of being somewhere in "b"}]:}{[\text{Probability of being somewhere in "a"}]} = \exp\left[-(F_b - F_a)/kT\right] \tag{5.1}$$

The free energy $F = E - TS$ comprises the energy E and the entropy S. I believe you know what energy is. As to temperature, your knowledge, I think, is rather *intuitive*, so we will return to it again later. The problem of entropy is still more complex, so let us discuss it once again.

In the simplest case of a particle in the container (as we discussed in Lecture 4), $S = k \ln V$, where V is the volume accessible to the particle.

In what sense are we incorporating this entropy into the free energy in the form of $-TS$? When considering $\exp[-F/kT]$ irrespective of E, we have got $\exp[-(-TS)/kT]\ \exp[-(-Tk \ln V)/kT] \equiv V$, that is just the accessible volume defining *the number of accessible states* of the particle in space. The greater the entropy, the larger this number, and the higher the probability that the particle is somewhere in this volume.

In the general case (when the accessible states of the particles are limited not only by the walls surrounding the volume V but also by collisions between the particles), the entropy S is given by

$$S = k \times [\text{logarithm of the number of accessible states of the studied particle}] \tag{5.2}$$

In molecular physics, biology and chemistry not a single particle but a mole (6.022×10^{23}) of particles are usually considered (6.022×10^{23} is called the *Avogadro's number*). Then the entropy of a mole is given by

$$S = R \times [\text{logarithm of the number of accessible states of one particle}] \tag{5.3}$$

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00005-3

where $R = k \times (6.022 \times 10^{23}\ \text{mol}^{-1})$. The only difference between k and R is that k refers to a single particle and R to a mole of particles.

Remark: In classical physics we have no "number of accessible states"; we only have "accessible volume of coordinates q and impulses p". However, this volume can be transferred into number of quantum states using the Heisenberg Uncertainty Principle ($\Delta q \Delta p \sim \hbar$, where Δq, Δp are the uncertainties in q and p, while $\hbar$, the Planck's constant, is the elementary volume of the coordinate-impulse space) (Landau and Lifshitz, 1977, 1980].

Strictly speaking, the value:

$$F = E - TS \tag{5.4}$$

is known as the Helmholtz free energy. It is easy to describe and convenient to calculate this value, since it refers to a system that (like the molecule we have just discussed) is enclosed in some fixed volume.

However, a normal experiment deals not with a fixed volume V but with a constant pressure P (eg, atmospheric pressure). In this case, as you may remember, we measure not a change in the energy E of the studied body but a change in its *enthalpy* $H = E + PV$. The change in enthalpy H includes, apart from the change in energy (E), the work against the external pressure P in changing the body volume V.

The value of PV will be negligible for all objects considered in these lectures, since we will deal with liquids or solids (where the volume per molecule is small) at a rather low (eg, atmospheric) pressure. Under these conditions, the value of PV is many times less than the thermal energy of the body.

Indeed, even for a gas (where the volume per molecule is particularly large), $PV \approx RT \times$ [number of moles] (remember the Clapeyron-Mendeleev law?). So, per mole of gas, the correction $PV = 1RT \approx 0.55/0.75$ kcal mol^{-1} at temperatures ranging from 0 to 100°C, that is, at $T = 273$ to 373 K. In other words, even in gases this value is commonly low compared with the magnitudes of the effects of interest, which usually amount to a few kilocalories per mole. Under a pressure close to atmospheric, for liquids and solids, the effect on H of correcting for PV is still hundreds or thousands of times less: here the volume of one mole is a minor fraction of a liter [$\approx 1/55$ L for H_2O, $\approx 1/10$ L for $(CH_2)_6$, etc.], while for a gas at a pressure of 1 atmosphere and room temperature this volume amounts to about 25 L.

That is why, later I will neglect the difference between H and E and refer to them both simply as "energy."

Similarly, there is little difference (for us) between *Helmholtz free energy* $F = E - TS$ and *Gibbs free energy*

$$G = H - TS = (E + PV) - TS = F + PV \tag{5.5}$$

Making no difference between G and F we will refer to them both simply as "free energy." However, it is worth keeping in mind that for processes occurring

in a fixed volume, we should use E and F, while for processes occurring under constant pressure, we should use H and G letters.

Next, you should remember that at a given temperature ($T = \text{const.}$), any system adopts the equilibrium *stable* state at the minimum $F = E - TS$ if the *volume* is fixed, and at the minimum $G = H - TS$ if the *pressure* is constant. (The number of particles in the system is assumed to be fixed, unless the opposite is defined.)

With a minor change in the state of the system, its free energy varies as:

$$\begin{aligned} &F \to F + dF = F + dE - TdS - SdT \\ &\text{or} \\ &G \to G + dG = G + dH - TdS - SdT \end{aligned} \tag{5.6}$$

This means that all possible rearrangements of the system about its stable state are described by the following equations characterizing the free energy minimum—as we know, a peculiarity of the point of minimum (and maximum) is that minor deviations from it result in almost no change in the function's value:

1. at $V = const$, the stable state (at a given T) is gained with F at a minimum, where

$$dF|_{V=\text{const}} = dE|_{V=\text{const}} - TdS|_{V=\text{const}} = 0 \tag{5.7}$$

(taking into account that $dT = 0$ at $T = \text{const.}$, that is, $SdT = 0$ in Eq. (5.6));

2. at $P = \text{const.}$, the stable state (at a given T) is gained with G at a minimum, where

$$dG|_{P=\text{const}} = dH|_{P=\text{const}} - TdS|_{P=\text{const}} = 0 \tag{5.8}$$

These equations yield *the thermodynamic definition of absolute temperature*:

$$T = \left[\frac{dE}{dS}\right]_{V=\text{const}} = \left[\frac{dH}{dS}\right]_{V=\text{const}} \tag{5.9}$$

I realize that this definition has been obtained rather formally, and its physical sense is not obvious. Therefore, I will return to it later.

Now let us consider the chemical potential. This quantity describes the thermodynamic characteristics of one molecule in a system, rather than those of the system as a whole.

If molecules are added to a system one by one *under constant pressure*, identical efforts are required for driving in each of them. (This is not so if we add particles one by one to a system having a constant volume—it takes more and more effort to drive more particles in: it is easy to inject one drop into a sealed bottle, but what about another? One more? Still more?) *Under constant pressure*, the volume will grow when we add particles, while the density of the

system and the intensity of interactions in its interior will remain unchanged. That is, *under constant pressure H* and *G are* proportional to N, the number of particles in the system (while at a constant volume, E and F are *not* proportional to N).

Thus, the thermodynamic state of a single molecule in a large homogeneous system is adequately described by the Gibbs free energy G divided by the number of molecules N,

$$\mu = \frac{G}{N} \tag{5.10}$$

where μ is known as the *chemical potential*, which can be also defined as the work spent to add one more particle to the system: $\mu = (dG/dN)_{T,P=\text{const}} = (dF/dN)_{T,V=\text{const}}$ (Landau and Lifshitz, 1980) (and since in liquids or solids $F \approx G$ at low pressures, here $\mu \approx F/N$). However, $G = N\mu$ if the system consists of identical molecules. If there are different types of molecules ($i = 1, 2, \ldots$) in the system, then $G = \Sigma_i N_i \mu_i$. *Note*: If N means not the number of molecules but, as usual in physical chemistry, the number of moles of molecules, then μ refers not to one molecule but to a mole of molecules.

The chemical potential, or, which is the same, the Gibbs free energy per molecule, will be of use later in this lecture for considering the distribution of molecules between phases. The thing is molecules pass from the phase where their chemical potential is higher to a phase where it is lower, thereby lowering the total free energy of the system and shifting it to equilibrium. And the equilibrium state is characterized by identical chemical potentials of molecules in both phases.

For further considerations, we will need two more equations.

First, the definition of heat capacity reflecting a temperature-dependent increase of energy:

$$C_P = \left[\frac{dH}{dT}\right]_{P=\text{const}} \tag{5.11}$$

(this is for a constant pressure; one can also calculate heat capacity at constant volume, but we do not need such details).

Second, the relationship between the entropy and the free energy:

$$S = -\left[\frac{dG}{dT}\right]_{P=\text{const}} = -\left[\frac{dF}{dT}\right]_{V=\text{const}} \tag{5.12}$$

Eq. (5.12) is one of the most important in thermodynamics. It results directly from the fact that the small increment of free energy $dG = d(H - TS) = dH - TdS - SdT$ (and $dF = d(E - TS) = dE - TdS - SdT$), whereas, in equilibrium, $dH - TdS = 0$ (and $dE - TdS = 0$) in accordance with the thermodynamic definition of temperature: $T = dH/dS = dE/dS$ (see Eq. 5.9).

Eq. (5.12) shows that the free energy has its minimum (or maximum) value at such temperature T, where $S(T) = 0$.

It is also useful to know that G/RT is at a minimum (maximum) when $H(T)=0$, since

$$\frac{d(G/RT)}{dT}=\frac{d(G/dT)}{RT}-\frac{G}{RT^2}=\frac{-S}{RT}-\frac{(H-TS)}{RT^2}=\frac{-H}{RT^2} \tag{5.13}$$

Remarks:

1. Using Eqs. (5.5) and (5.12), one can show that $H=G+TS=G-T\,(dG/dT)$, and therefore C_P (see Eq. 5.11) can be obtained in the form:

$$C_P=-T\left[\frac{d^2G}{dT^2}\right]_{P=\text{const}} \tag{5.14}$$

2. We are never interested in values of energy, entropy, and free energy themselves. We are interested only in *changes* of these values. Indeed, speaking on energy: we can count off the gravitational energy of particles from the sea level, from the floor level, from the center of the Earth, etc. The values will be all different, but this does not matter. The only important thing is the *difference* between gravitational energies of a particle in two states. Just the same, when we define the particle's entropy as k ln V, we can measure the volume V in liters, in cubic feet, etc. The values will be all different, but this does not matter: the only important thing is the *difference* between particle's entropies in two volumes (V and V'. And this difference, $\Delta S=k$ ln(V/V'), is independent of the unit of measurements.

With this introduction, we can now go on to discuss hydrophobicity and water as a specific solvent that creates it.

First of all, let us consider the so-called *hydrophobic effect*.

"Hydrophobicity" is "fear of water." Who is "afraid" of water?—all nonpolar molecules such as inert gases (argon, xenon), hydrogen, and all purely hydrocarbon molecules (methane, ethane, benzene, cyclohexane, etc.). We will focus on hydrocarbons in water, since proteins have many hydrocarbon side chains. It is these water-fearing and water-escaping side chains that form the *hydrophobic core of a protein globule* (Bresler and Talmud, 1944; Kauzmann, 1959).

So, what does hydrophobicity mean in terms of experiment?

Methane's (CH_4) concentration in water is about an order *lower* than in gas above this water at a temperature of 20–40°C (the exact value depends on temperature). For H_2 and for propane $CH_3CH_2CH_3$, the difference is nearly the same. All of them are water-fearing (phobic) molecules: they are *more numerous* in a vapor than in the water. In contrast, ethanol CH_3CH_2OH and water are easily miscible, and as you know, their separation is quite laborious, but the ethanol molecule is polar (its polar O–H group is capable of H-bonding). As to the purely nonpolar molecules, they would prefer a vacuum (vapor is nearly a vacuum) rather than water. And this happens *in spite* of van der Waals attraction between *any* molecules, even H_2 or CH_4 and water.

Let us consider (using one example shown in Fig. 5.1) some typical thermodynamic effects for hydrocarbons in water. The thermodynamic parameters presented in Fig. 5.1 (free energies, energies, and entropies of transfer from one phase to another) resulted from experimental studies of the equilibrium distribution of molecules of cyclohexane, $(CH_2)_6$, among three phases: vapor, aqueous solution, and liquid cyclohexane.

If molecules can go from one phase to another, their chemical potentials μ, in equilibrium, are equal in all the phases; otherwise, molecules will start to leave the phase with a higher chemical potential and go to the phase with a lower chemical potential: this will decrease the free energy.

Chemical potential for a given type of molecules in a given phase (where they have concentration $X \equiv N/V$) may be presented as

$$m = G^{\text{int}} + RT \ln(X), \tag{5.15}$$

where G^{int} is the "mean force potential" acting upon these molecules in this phase (ie, the free energy of interactions of the molecules in question with their surrounding).

For equilibrium distribution of given molecules between two phases (nonpolar liquid and aqueous solution), $\mu_{\text{in liquid}} = \mu_{\text{in aqueous solution}}$. Thus, the

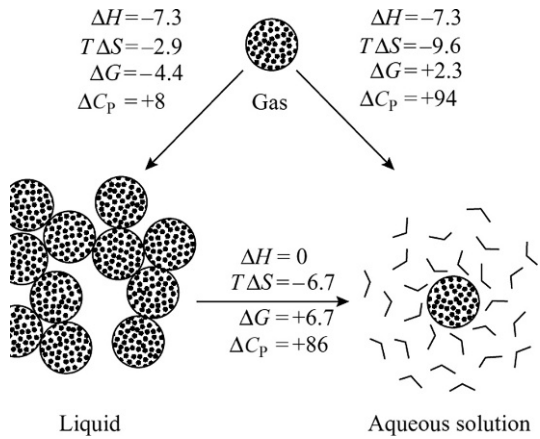


FIG. 5.1 The thermodynamics of transfer of a typical nonpolar molecule, cyclohexane $(CH_2)_6$, from vapor (top) to water (right), and from liquid cyclohexane (left) to water. The numerical values are for 25°C (ie, $T \approx 300$ K, $RT \approx 0.6$ kcal mol^{-1}). ΔH (as with energy, it is measured in "kilocalories per mole") is the interaction enthalpy change per mole of molecules; ΔS is the corresponding interaction entropy change ($T\Delta S$, the contribution of the entropy change to the free energy, is also measured in kilocalories per mole); $\Delta G = \Delta H - T\Delta S$ (kcal mol^{-1}) is the change in Gibbs free energy of interactions per mole of transferred molecules; ΔC_P [cal mol^{-1} K^{-1}] is the change in heat capacity per mole of transferred molecules. All values are recalculated using Eqs. (5.16)–(5.18) and experimental values from reference books. *(Adapted from Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W. H. Freeman & Co., New York (Chapter 4).)*

difference in the free energy of interactions upon transfer of a mole of molecules from the nonpolar liquid to water is defined as:

$$\Delta G_{\text{liquid}\rightarrow\text{aqueous solution}} \equiv G^{\text{int}}_{\text{in aqueous solution}} - G^{\text{int}}_{\text{in liquid}} = -RT\ \ln\left(X_{\text{aq}}/X_{\text{liq}}\right) \tag{5.16}$$

Here X_{aq} and X_{liq} are equilibrium concentrations of studied molecules in the aqueous solution and in the nonpolar liquid that contacts with the former.

The value $\Delta G_{\text{liquid}\rightarrow\text{aqueous solution}}$ is a difference between the mean force potentials affecting our molecule in water and in the nonpolar liquid. The mean force potential is created by all interactions of our molecule with its molecular surrounding; it includes both energy and entropy terms that arise from these interactions. Since concentration is the number of molecules in a given volume, Eq. (5.16) follows from the Boltzmann distribution over two phases with different values of the mean force potential.

At 25°C and low ($\sim$1 atm. or less) pressure of the air, $X_{\text{liq}} \approx 9.25\ \text{mol}^{-1}$ for pure liquid $(CH_2)_6$ and $X_{\text{aq}} \approx 0.0001\ \text{mol}^{-1}$ for its saturated solution in water. Accordingly, $\Delta G_{\text{liquid}\rightarrow\text{aqueous solution}} = +6.7\ \text{kcal mol}^{-1}$.

The free energy of transfer of our molecule from gas to the nonpolar liquid $(CH_2)_6$ and to the aqueous solution can be defined in the same way:

$$\Delta G_{\text{gas}\rightarrow\text{liquid}} = -RT\ \ln\left(X_{\text{liq}}/X_{\text{gas}}\right) \tag{5.17}$$

$$\Delta G_{\text{gas}\rightarrow\text{aqueous solution}} = -RT\ \ln\left(X_{\text{aq}}/X_{\text{gas}}\right), \tag{5.18}$$

where X_{gas} is the equilibrium concentration of our molecules in gas above the liquid(s). The pressure of the saturated $(CH_2)_6$ vapor at 25°C is about 0.05 atm, which means that $X_{\text{gas}} = 0.002\ \text{mol}^{-1}$ (ie, X_{gas} is 50 times higher than X_{aq}). The resulting ΔG values are presented in Fig. 5.1. *Note*: the experimentally measured value of $X_{\text{aq}}/X_{\text{gas}}$ at $X_{\text{gas}} \rightarrow 0$ is called "Henry's law constant" ($k_{\text{H,cc}}$).

Since the mean force potential of interactions affecting a molecule in a rarefied gas ($G_{\text{in gas}}$) is virtually zero, $\Delta G_{\text{gas}\rightarrow\text{liquid}} = G_{\text{in liquid}} - G_{\text{in gas}}$ is very close to $G_{\text{in liquid}}$ that is the mean force potential of interactions of the molecule in the nonpolar liquid; and the mean force potentials of the molecule's interactions in water, $G_{\text{in aqueous solution}}$, equals to $G_{\text{gas}\rightarrow\text{aqueous solution}}$.

When ΔG and its temperature dependence is known, the values of ΔS, ΔH, ΔC_P are derived from this dependence according to Eqs. (5.12)–(5.14).

Fig. 5.1 shows that the *energy* of attraction of $(CH_2)_6$ molecules to water is as high as that to liquid cyclohexane ($\Delta H = -7.3\ \text{kcal mol}^{-1}$), but $(CH_2)_6$ molecules do not want to go into water, though they go into liquid cyclohexane readily. As is clear from Fig. 5.1, it is entropy that causes this hydrophobicity; it becomes too low when a cyclohexane molecule comes into water.

Why are nonpolar molecules like CH_4 or $(CH_2)_6$ hydrophobic?

The reason is that, unlike water molecules, nonpolar H_2 or Ar, CH_4 or $(CH_2)_6$ are incapable of H-bonding. This is confirmed by the well-known fact that polar ethanol molecule, CH_3CH_2OH (which consists mainly of

hydrocarbon groups, such as $(CH_2)_6$, but has a –OH group and therefore is capable of H-bonding, like H_2O) is not hydrophobic.

A naive suggestion could be made that upon coming into water, CH_4 or $(CH_2)_6$ molecules disrupt H-bonding in water. But it is not that simple. If it were so, the solution energy would have *increased* drastically with incoming $(CH_2)_6$, while in fact it actually *decreases* (see Fig. 5.1) by 8 kcal mol^{-1} of incoming $(CH_2)_6$. Actually, instead of increasing energy, we have *decreasing entropy* (and this decrease is substantial: $T\Delta S = -9.6$ kcal mol^{-1}).

This entropy decrease prevents cyclohexane from dissolving in water. The free energy $G = H - TS$ increases not only with increasing energy H but also with *decreasing* entropy S (S contributes to G in the form of $-TS$). Thus, a large decrease of the entropy S, even with a simultaneous decrease of energy H, results in increasing free energy G, and hence (see Eq. (5.1)) in a *decreasing probability* of molecules remaining in the current state (or more simply, in their decreasing concentration in this state).

Now the main physical question arises as to why the entropy of water molecules decreases as a result of their contact with a nonpolar surface.

It decreases because an H_2O molecule must not point at a hydrophobic surface with its H atom. Otherwise, its hydrogen bonds will be lost (see Fig. 5.2; as you may remember, H-bonds are orientation-dependent and emerge only when an O–H group is directed towards the O atom of another water molecule).

Let me remind you that H_2O molecules are almost fully hydrogen-bonded in water, so it is impossible to sacrifice some H-bonds without a great loss of the free energy. To avoid the loss of H-bonds (ie, to avoid O–H groups being directed towards the hydrophobic surface), water molecules seek favorable positions (see the upper molecule in Fig. 5.2) and partially freeze their thermal motions. Thereby, they preserve their valuable H-bonds at the expense of some

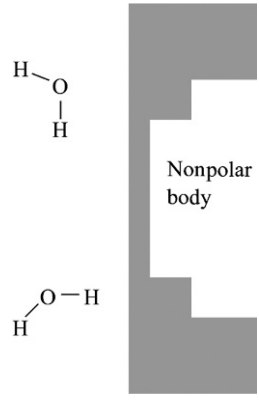


FIG. 5.2 Water molecules near the surface of a nonpolar body. The upper molecule can make all its H-bonds, but this "favorable" position about the body surface is restricted and therefore entropy-expensive. The lower molecule loses one energy-expensive H-bond to water because its O–H group is directed towards the nonpolar obstacle.

of their entropy. From the data presented in Fig. 5.1 and the number of waters surrounding $(CH_2)_6$, we can estimate that the entropy-driven loss of free energy is about 0.2 kcal per mole of mean-surface waters, which is an order of magnitude less than the price of lost H-bonds would be.

Low temperature (up to 20°C, and for some hydrocarbons, up to 60°C) even allows the hydrogen bonds near a hydrophobic surface (see Fig. 5.3) to gain a little energy (see Fig. 5.4; because now these bonds are less damaged by movements of half-frozen near-surface waters), but this gain does not fully compensate for the entropy loss resulting from freezing the near-surface waters.

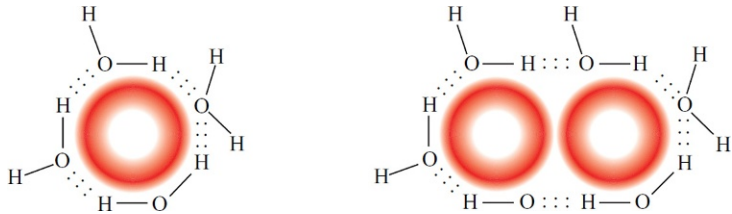


FIG. 5.3 Irregular packing of H-bonded waters around a nonpolar molecule (left) and around a pair of such molecules. A hydrophobic bond is formed in the latter case.

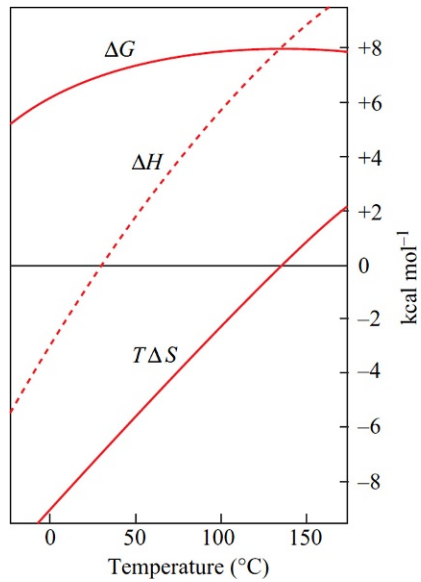


FIG. 5.4 The thermodynamics of transferring a typical nonpolar molecule, pentane C_5H_{12}, from liquid pentane to water at various temperatures. The transfer free energy, $\Delta G = \Delta H - T\,\Delta S$, and its enthalpic ($\Delta H$) and entropic ($T\Delta S$) components are measured in kilocalories per mole of transferred molecules. ΔG is at a maximum when $\Delta S = 0$; the proportion of pentane distribution between the water and the liquid pentane phases (which is proportional to $\exp(-\Delta G/RT)$) is at a minimum when $\Delta H = 0$. *(Adapted from Privalov, P.L., Gill, S.J., 1988. Stability of protein structure and hydrophobic interaction. Adv. Protein Chem. 39, 191–234.)*

Note that again the net effect is entropic rather than energetic in nature *just because* the energy of H-bonds is extremely high: since it is so, waters would prefer to become frozen (although this is also thermodynamically bad) and sacrifice a part of their freedom (entropy) than to lose the large energy of a hydrogen bond.

I would like to emphasize that the resultant entropic effect on the free energy value is of *the same sign* as the energetic effect expected by naiveté, but *less* in magnitude.

The suggestion that waters close to a nonpolar surface are, in a way, frozen, is additionally supported by the anomalously high heat capacity of cyclohexane (and other hydrocarbons) in water. The excessive heat capacity of a $(CH_2)_6$ molecule in aqueous surrounding is 10 times as high as that amidst its fellow cyclohexanes. To be more exact, a high heat capacity in water is typical not of the hydrocarbon itself but of its ice-like water shell; with increasing temperature this "iceberg" tends to melt out, and this explains the anomalous heat capacity.

In considering frozen hydrogen-bonded surface waters, bear in mind that their relative orientation *differs* from that in normal ice. In ice, water molecules must have regular space positions because they have to form a huge three-dimensional crystal. At the surface they can adopt any position they like, provided that it is favorable for H-bonding. The water molecules do not observe translational symmetry of the three-dimensional lattice of ice because the resultant "microiceberg" is not going to grow infinitely: it only tends to coat the introduced hydrophobic molecule (Fig. 5.3) or a group of such molecules.

In the latter case, these hydrophobic molecules form a "hydrophobic bond."

Inner voice: There is "hydrophobic bond" and there is "hydrogen bond." Both bonds are hidden from water—so, what's the difference between them?
Lecturer: Hydrogen bond is a bond between differently charged atoms of polar groups, say, $N^-H^+:::O^-C^+$, and its strength is *decreased* by water environment (as the strength of electrostatic bond, which will be considered in the next lecture); therefore, it is not called "hydrophobic." Hydrophobic bond is a bond between nonpolar ("hydrophobic") groups, say, $(CH_2)_6:::(CH_2)_6$ or $CH_4:::C_5H_{12}$, and its strength is provided by ordering of waters surrounding these groups. (That is, hydrophobic bond is stronger than a simple van der Waals bond between the same groups.)

The so-called *clathrates* represent an extreme ordering of waters caused by hydrophobic molecules. Clathrates are crystals built up by water and nonpolar molecules (Pauling, 1970).

From your chemistry course, you may know that they are far less stable than the crystalline hydrates built up by water and polar molecules. Clathrates emerge only at low temperatures (about 0°C or less) and high pressures, which cause many molecules of nonpolar gas to penetrate into water. In clathrates, as in ice, water molecules have their hydrogen bonds saturated, although their geometry is different from that in normal ice. In the resultant crystal, quasi-

ice keeps nonpolar molecules in its pores. Incidentally, clathrates are thought to contain more natural gas than ordinary gas fields, and gas production from clathrates (from a great depth where high pressure ensures their existence) is perhaps a project of tomorrow.

The hydrophobic effect is rather temperature-dependent (Fig. 5.4). The temperature affects ΔG, but even more, it affects the magnitude (and even sign) of its constituents, ΔH and ΔS.

As the temperature is increased, the surface hydrogen bonds tend to melt out. Up to $\approx 140°C$, this is accompanied by an increasing hydrophobic effect, since the thermodynamically unfavorable ordering of surface waters persists, while favorable hydrogen bonds are destroyed.

While at low and room temperature the hydrophobic effect results from entropy only, at elevated temperatures, the energy of the lost near-surface hydrogen bonds becomes more and more important. But ΔG keeps increasing until $\Delta S < 0$, ie, up to $\approx 140°C$ (see Fig. 5.4).

However, at still higher temperatures, there are too many disrupted H-bonds in water (which now remains liquid only under high pressure); as a result, the hydrophobic surface interferes with hydrogen bonding less and less, and the hydrophobic effect begins to diminish.

Fig. 5.4 illustrates the transfer of a nonpolar molecule to water from a nonpolar solvent rather than from a vapor. This is done deliberately: we are going to consider the hydrophobic effect in proteins where amino acid residues are transferred from water to the protein core, which is similar to a nonpolar solvent rather than to a vapor.

The hydrophobicity of amino acids will be in focus later but now it would be useful to consider Table 5.1; it contains some characteristics measured at room temperature for nonpolar groups similar to protein ones.

It is immediately apparent from Table 5.1 that ΔG (unlike ΔH and ΔS) increases with increasing size of a hydrophobic molecule. How exactly does it increase? A more detailed analysis of a variety of nonpolar molecules shows that their hydrophobic free energy ΔG increases almost proportionally to the accessible surface area of the nonpolar molecules.

The physical sense and mode of construction of the accessible surface is demonstrated in Fig. 5.5 (see also Fig. 5.3).

The hydrophobic free energy is about +0.02–+0.025 kcal mol^{-1} $Å^{-2}$ for the accessible nonpolar area of a molecule transferred from a nonpolar solvent to water (Chothia, 1974). Specifically, for benzene, the accessible area is about 200 $Å^2$, and $\Delta G \approx 4.6$ kcal mol^{-1}; for cyclohexane, the accessible area is about 300 $Å^2$, and $\Delta G \approx 6.7$ kcal mol^{-1}.

The same regularity is observed (Chothia, 1974) for hydrophobic amino acid residues (Fig. 5.6).

The hydrophobicities of amino acids are derived experimentally from equilibrium distributions of amino acids between water and a nonpolar or slightly polar solvent; the latter is usually a high-molecular-weight alcohol (eg, octanol)

TABLE 5.1 Typical Thermodynamic Parameters of Hydrophobic Group Transfer from a Nonpolar Liquid to an Aqueous Solution at 25°C

Molecule	Transfer from → to	ΔG (kcal mol^{-1})	ΔH (kcal mol^{-1})	$T\Delta S$ (kcal mol^{-1})	C_P (kcal mol^{-1} K^{-1})
Ethane $(CH_3)_2$ (compare with Ala side group: $-CH_3$)	Benzene → water	+3.6	−2.2	−5.8	+59
	CCl_4 → water	+3.8	−1.8	−5.4	+59
Benzene C_6H_6 (compare with Phe side group: $C_6H_5-CH_2-$	Benzene → water	+4.6	+0.5	−4.1	+54
Toluene $C_6H_5-CH_3$ (compare with Pheside group)	Toluene → water	+5.4	+0.4	−5.8	+63

Values taken from Tanford (1980).

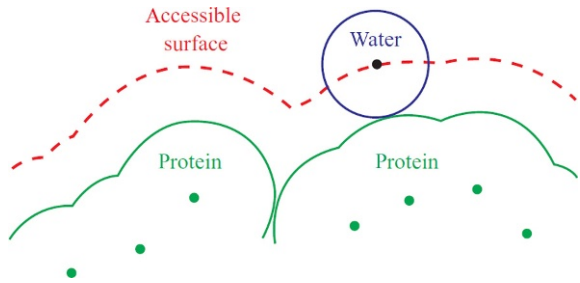


FIG. 5.5 The "accessible surface" of a molecule in water. Dots indicate the centers of the molecule's atoms exposed to water; the solid line denotes their van der Waals envelopes. A water molecule is shown as a sphere of radius 1.4 Å. The "accessible" (to water) surface is defined as the area described by the center of a 1.4 Å sphere that rolls over the van der Waals envelope of the protein. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979. Principles of Protein Structure. Springer, New York (Chapter 1).)*

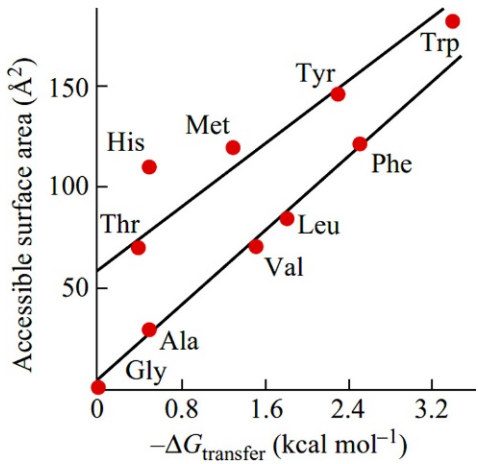


FIG. 5.6 The accessible surface area of amino acid side chains and their hydrophobicity. (The accessible surface of a side chain X is equal to the accessible surface of amino acid X minus that of Gly having no side chain; the hydrophobicity of a side chain X is equal to the experimentally measured hydrophobicity of amino acid X minus that of Gly.) The side chains of Ala, Val, Leu, Phe consist of hydrocarbons only. Those of Thr and Tyr additionally have one OH-group each, Met–SH-group, Tpr–NH-group, and His–N-atom and NH-group; therefore, their accessible *non-polar* area is smaller than their total accessible surface. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979. Principles of Protein Structure. Springer, New York (Chapter 1), with minor modifications.)*

or dioxane. These experiments are far from simple because some amino acids are hardly soluble in water, while others are hardly soluble in organic solvents (for instance, polar amino acids are virtually insoluble in purely nonpolar cyclohexane or benzene). Therefore, a nonpolar solvent is replaced by a specially selected slightly polar one (such as octanol, which is satisfactory for both strongly and weakly polar amino acids); other tricks are used too (Fauchére

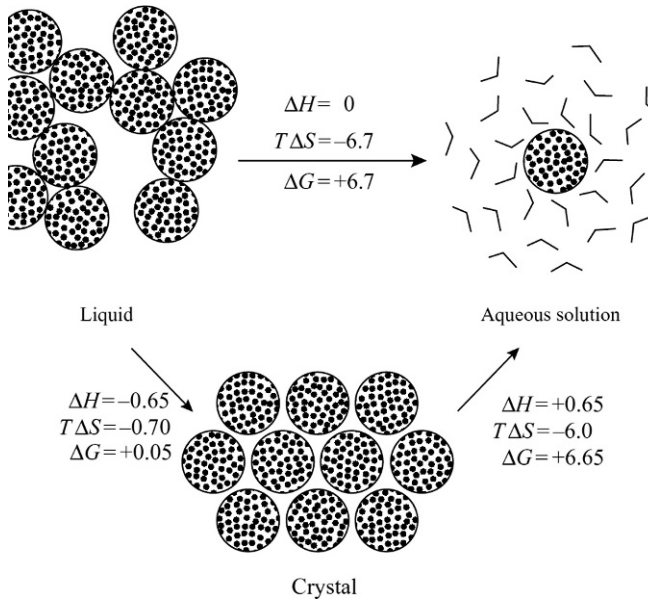


FIG. 5.7 The thermodynamics of transfer of a typical nonpolar molecule, cyclohexane, from the liquid to the solid phase and to aqueous solution. The numerical values are for 25°C. H, $T\Delta S$ and ΔG are measured in kilocalories per mole. All values are recalculated using Eqs. (5.16)–(5.18) and experimental data from reference books. *(Adapted from Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W. H. Freeman & Co., New York (Chapter 4).)*

and Pliska, 1983). According to the various solvents used, the results differ, specifically for charged and strongly polar amino acids. However, the qualitative agreement of these results is not bad.

The hydrophobic effect is smaller for the side chains with polar atoms than for completely nonpolar groups (provided their total accessible surface area is the same). However, if we consider only the nonpolar-atom-produced portion of their accessible surface area (ie, the total accessible surface area minus approximately 50 $Å^2$ for each polar atomic group), essentially the same surface dependence of hydrophobicity can be revealed for all the groups (see Fig. 5.6).

The hydrophobic effect is of major importance in maintaining the stability of the protein structure. It is this effect that is responsible for the compact globule formation of the protein chain (Bresler and Talmud, 1944; Kauzmann, 1959). As seen from Fig. 5.7, the free energy of transfer of a hydrophobic group from water to hydrophobic media is high and amounts to a few kilocalories per mole, while the free energy of hardening of the nonpolar liquid is close to zero at physiological temperatures. To be more accurate, I should say that the free energy of hardening of comparatively small hydrocarbons (see cyclohexane, shown in Fig. 5.7) is actually positive, which prevents their hardening at room temperature. The hardening is prevented by entropy of rotations and movements

of a molecule in the liquid, where each molecule is more or less free, in contrast to a solid in which the crystal lattice keeps the molecules fixed. However, the entropy of these motions of a molecule does not depend on its size, while the enthalpy of a molecule increases with increasing number of intermolecular contacts, that is, with the size of the molecular surface. In a side chain, the entropy of motions is lower, since amino acid residues are linked into a chain and cannot move freely, and this facilitates hardening.

However, even if the entire entropic component of cyclohexane crystallization, $\Delta G_{\text{liquid}\rightarrow\text{crystal}}$, was neglected (ie, if we assume that $\Delta G_{\text{liquid}\rightarrow\text{crystal}} \approx \Delta H_{\text{liquid}\rightarrow\text{crystal}} = -0.65$ kcal mol^{-1}), the thermodynamic effect of crystallization would be much weaker than the hydrophobic effect condensing the water-dissolved cyclohexane molecules into a liquid drop ($\Delta G_{\text{aqueous solution}\rightarrow\text{liquid}} = -6.7$ kcal mol^{-1}).

Thus, loosely speaking, the hydrophobic effect is responsible for approximately 90% of the effort required to make a compact globule. However, by itself it cannot provide a native solid protein. It creates only the molten globule that is yet to be discussed. The hardening of a protein, like that of all compounds, results from van der Waals interactions, as well as from hydrogen and ionic bonds, which are far more specific and more sensitive to peculiarities in the atomic structure than simple hydrophobicity (which is, actually, an effect operating in water). But what they perform is the final "polishing" of the native protein, whereas the bulk of the basic work is done by the hydrophobic effect.

REFERENCES

Bresler, S.E., Talmud, D.L., 1944. On the nature of globular proteins. II. Some consequences of a new hypothesis. Dokl. Akad. Nauk SSSR (in Russian) 43, 326–330; 367–369.

Cantor, C.R., Schimmel, P.R., 1980. Biophysical chemistry. W.H. Freeman & Co, New York (Part 1, Chapters 2, 5).

Chothia, C., 1974. Hydrophobic bonding and accessible surface area in proteins. Nature 248, 338–339.

Fauchére, J.L., Pliska, V., 1983. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur. J. Med. Chem. Chim. Ther. 18, 369–375.

Kauzmann, W., 1959. Some factors in interpretation of protein denaturation. Adv. Protein Chem. 14, 1–63.

Landau, L.D., Lifshitz, E.M., 1977. Quantum Mechanics. A Course of Theoretical Physics, vol. 3. Pergamon Press, Oxford, New York (Chapters 1, 2).

Landau, L.D., Lifshitz, E.M., 1980. Statistical Physics, third ed. A Course of Theoretical Physics, vol. 5. Elsevier, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo Sections 7, 14, 15, 24.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co, New York (Chapter 2).

Pauling, L., 1970. General Chemistry. W.H. Freeman & Co, New York (Chapters 10, 12).

Tanford, C., 1980. The Hydrophobic Effect, second ed. Wiley-Interscience, New York.

This page intentionally left blank

Lecture 6

In this lecture, we discuss electrostatic interactions and, specifically, their features induced by the protein globule and its aqueous environment.

It may seem that there is nothing to discuss; undoubtedly, you remember that in a medium with permittivity (dielectric constant) ε the charge q_1 creates an electric field whose potential at the distance r is:

$$\varphi = \frac{q_1}{\varepsilon r} \tag{6.1}$$

and q_1 interacts with the charge q_2 at this distance with the energy

$$U = \varphi q_2 = \frac{q_1 q_2}{\varepsilon r} \tag{6.2}$$

You may also remember that in a vacuum (or air) $\varepsilon = 1$, in water ε is close to 80, and in media such as plastics (and dry protein as well) ε is somewhere between 2 and 4.

> *Inner voice:* All this is true, or, better to say, almost true. First, strictly speaking, U is not the energy but the *free* energy that tends to a minimum when our charges are in some medium instead of a vacuum; and U depends on permittivity ε of the medium and hence varies with temperature together with ε; this means that U contains an entropic part.
> *Lecturer:* This comment is absolutely right, but inasmuch as we are considering how charge interactions affect protein stability, these are minor items, a matter of purism. For the sake of simplicity, let me use the term "energy" for a while…

What is more important, Eq. (6.2) is valid only for homogeneous media. And when studying charge interactions in proteins, we are dealing with a most heterogeneous medium. The permittivity of a protein itself, as with that of plastics, is not high and amounts to about 2–4, while that of water is 80. And the charged groups of the protein are mostly located on its surface, close to the water (we will see why later). What value of ε then should be chosen to make estimates of electrostatic interactions in the protein? If we take $\varepsilon \approx 80$, the energy of interaction between two elementary (proton) charges at a 3 Å distance will be approximately 1.5 kcal mol^{-1}; with $\varepsilon \approx 3$, this energy will be about 40 kcal mol^{-1}. The difference is too large: the additional 40 kcal mol^{-1} can destroy any protein structure. (It has been already mentioned that the typical "reserve of stability" of a protein structure—the difference in free energies between the native and denatured states of the protein—is about 10 kcal mol^{-1}; any effect exceeding this value causes an "explosion" of the structure.)

The other problem is as follows: Eqs. (6.1) and (6.2) are valid when the distance r between charges much exceeds the size of surrounding molecules.

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00006-5

However, in proteins, charges are often in immediate contact with as little as 3–4 Å distance between them, which does not allow even a water molecule, not to mention a side chain, to take an intervening position. How can we estimate electrostatic interactions in this case? Should we assume that $\varepsilon = 1$, as in a vacuum? Or should we take $\varepsilon = 80$? Or rather…?

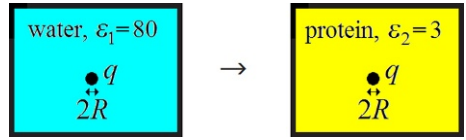
A brief philosophical digression. Why are these rough estimates wanted at all? Indeed, it is often believed that with the powerful computers available now, one can input "all as it is in reality": water molecules, the coordinates of protein atoms including the coordinates of the charges, the temperature (ie, the energy of thermal motion) and obtain "the precise result." As a matter of fact, this is a rather Utopian picture. The calculation—I mean the detailed one (made using the so-called "molecular dynamics")—will take hours or days because you will have to follow both thermal motions and polarization of many thousands of interacting atoms (and, by the way, will not be absolutely accurate either; if nothing else, remember that atoms are "nonspherical" owing to their p-orbitals and other quantum effects; however, this fact is ignored by the interaction potentials (Levitt et al., 1995; Halgren, 1995; Jorgensen et al., 1996; MacKerell et al., 1998; Wang et al., 2004; Donchev et al., 2008; Pereyaslavets, Finkelstein) used in molecular dynamics and conformational analysis. And what really interests you is most likely a simple quick estimate, such as whether it is possible to introduce a charge into the protein at this or that site without a risk of protein explosion. My aim is to teach you how to make such estimates.

First of all, let us estimate the change in energy of a charge upon its transfer from water ($\varepsilon \approx 80$) into the middle of the protein (where $\varepsilon \approx 3$). For the time being, let us use classical electrostatics: consider water and protein as continuous media and disregard their corpuscular (ie, atomic) structures; or rather, let us postpone considering those details.

According to classical electrostatics, a sphere of charge q and radius R in a medium of permittivity ε has the energy:

$$U = \frac{q_1^2}{2\varepsilon R} \tag{6.3}$$

This expression directly follows from Eq. (6.2): when we charge up the sphere (from zero to $q1$) by bringing small charges dq onto its surface, each small charge dq increases the energy of the sphere by $dU = q\,dq/\varepsilon R$ (according to Eq. (6.2)), and the integral of $q\,dq$ from 0 to q_1 is $q_1^2/2$.



$$\Delta U_{1\to 2} = \frac{q^2}{2\varepsilon_2 R} - \frac{q^2}{2\varepsilon_1 R}$$

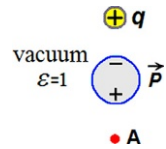
If the radius of the charged atom is about 1.5 Å, then its (free) energy is nearly 1.5 kcal mol^{-1} at $\varepsilon \approx 80$ (in water) and nearly 40 kcal mol^{-1} at $\varepsilon \approx 3$ (in protein). This great difference explains why inside the protein, in contrast to its surface, charged groups are virtually absent (it is easy to estimate that even immersion of a close pair of oppositely charged particles into the medium with $\varepsilon \approx 3$ increases the free energy by about the same magnitude of 40 kcal mol^{-1}). Therefore, an ionizable group is virtually always uncharged when it is deeply involved in the protein globule (Nelson and Cox, 2012): that is, a positively charged side-group donates its surplus H^+ to water, and a negatively charged side-group takes its missing H^+ from water. True, this discharging costs some additional free energy—but "only" a few kcal mol^{-1} (as you will see in Lecture 10), and not a few dozen kcal mol^{-1}. To be more exact, there are some charges in the interior of a protein; but these are almost always functional, and the protein has to put up with their presence just to keep functioning.

Now let us learn how to estimate the interaction of charges taking into account the interface between the protein ($\varepsilon \approx 3$) and water ($\varepsilon \approx 80$).

To begin with, let us consider a simple "problem of a charge and one dipole."

It is well known that increased (as compared with a vacuum) permittivity ε of the medium is created by the medium molecules, that is, dipoles oriented or polarized by an electric field. As a result, the medium $\varepsilon > 1$, and the field is *reduced* (see Eq. 6.1).

Suppose we have a charge $+q$, and our "medium" is a vacuum ($\varepsilon = 1$), and only *one* dipole is situated between the charge and the point A where the field potential is measured:



(the dipole $\vec{P}$ is naturally oriented along the field, that is, its "−" looks at our charge $+q$).

Question: Is the field at the point A reduced or increased by the dipole as compared to the 'vacuum case?

Answer: However odd and counterintuitive it might seem, the dipole *increases* the field at the point A! (Counterintuitive—because the medium filled with many such dipoles *decreases* the field, while one, the central and thus seemingly the most important dipole, does just the *opposite*!)

However, the ***answer*** was obtained very simply: The dipole "minus" is directed at our charge $+q$, while its "plus" looks in the opposite direction, that is, towards the point A; and because this "plus" is closer to A than "minus," its impact at the point A is *stronger* than that of the "minus," thereby *increasing* the field created by our $+q$ at the point A.

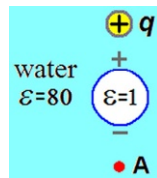
So, a single molecule located between the charge and the point where the field is measured reinforces the field, while the many molecules distributed over the entire space weaken the field! (Problem 6.1 shows that the field is reinforced by those molecules that are inside the sphere whose poles are the charge $+q$ and the observation point A, and weakened by molecules that are outside this sphere).

Note two important facts:

1. The same dipole $\vec{P}$ that enhances the potential at the point A, weakens the potential near the charge $+q$.
2. Dipole $\vec{P}$ is attracted to the charge $+q$, and *vice versa*.

We now consider a problem that is complementary to the first, and almost as simple: the "problem of a charge and a bubble."

Suppose that the "medium" is water (with $\varepsilon = 80$), and a vacuum bubble is located between the charge $+q$ and the point A where the field is measured:



Question: Is the field at the point A reduced or increased by the bubble as compared with the "pure water case"? *It may seem* that the field in A should be strengthened, because it goes from $+q$ to A through a space with a lower (on average) permittivity than that of pure water. But in fact…

Answer: Oddly, the field in A is reduced! Since the vacuum bubble is *not* polarized, and water is polarized, waters turn their "minuses" towards $+q$, and their "pluses" away from $+q$ (see Fig. 6.1). As a result, a positive induced charge arises at that surface of the bubble which faces $+q$, and a negative induced charge arises on the opposite surface of the bubble facing the point A (the sum of these induced charges is zero, according to a theorem in electrostatics given, eg, in Landau et al., 1984). The induced "minus," which is closer to A, weakens the potential created by $+q$ at the point A in pure water in the absence of a vacuum bubble.

Note two important effects opposed to those in the previous "problem of a charge and one dipole":

1. The same bubble that weakens the potential at the point A enhances the potential near the charge $+q$.
2. The bubble is repelled by the charge $+q$, and the charge $+q$ by the bubble.

Similar is the behaviour of the field of the charge positioned in water near the surface of a protein (Fig. 6.1). Here, polarization of the protein (with $\varepsilon \approx 3$) can be neglected, to the first approximation, as compared with that of water (with $\varepsilon \approx 80$).

Fig. 6.1 shows how polar water molecules are oriented around the protein and the charge $\oplus$. They are oriented in conformity with the field: their "$-$" are directed mostly at "our" charge $\oplus$, while their "+" are directed oppositely. This results, first, in partial compensation of the charge $\oplus$ by adjacent "$-$" of waters; this is trivial and simply causes a partial compensation of $\oplus$ and, as a result, a large water permittivity ε_1. But, second, this also produces the event that is of interest to us. Namely, the water "pluses," trying to be directed away from $\oplus$, have to turn to the protein side facing the charge $\oplus$, and this polarization induces a positive charge at this side of the water/protein interface, while the water "minuses" come to the protein on the other side. This induces there an opposite (negative) polarized charge (as mentioned above, polarization of the protein itself can be neglected, since $\varepsilon_{\text{protein}} \ll \varepsilon_{\text{water}}$).

As a result, on the $\oplus$-facing protein side, the field potential becomes higher compared with what would have been in the absence of the protein: here, the induced "pluses" add to the potential of $\oplus$ (and the induced "minuses" are far away and their effect is minor). That is why here $\varepsilon_{\text{eff}} \approx 40$ (see Problems 6.2–6.5 for the simplest, and Landau et al. (1984) and Finkel'shtein (1977) for more complicated cases).

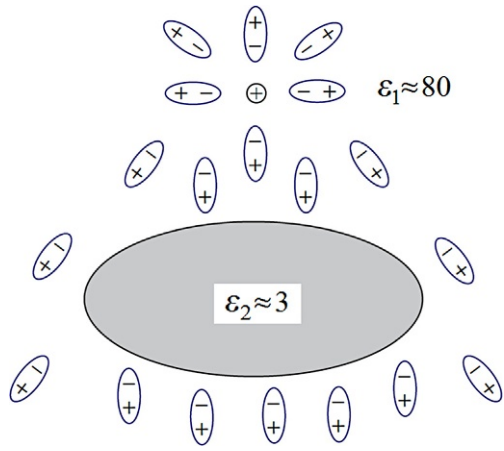


FIG. 6.1 The orientation of water molecules (shown as dipoles $(-+)$) around the protein (gray) and the charge $\oplus$ (which is shown as positive for the sake of simplicity only).

At the same time, on the ⊕-opposite side of the protein, the field potential becomes *lower* than it would have been in the absence of the protein: here, the potential of the charge ⊕ is diminished by the opposite in sign potential of "minuses" induced at this side of the protein/water interface (and the induced "pluses" are far away and their effect is minor). And since here (on the ⊕-opposite side of the protein), the field potential is *lower* than it would have been in the absence of the protein, ε_{eff} is *higher* than $\varepsilon_1 = 80$ (the value in the absence of the protein; see Landau et al., 1984 and Finkel'shtein, 1977 for calculations).

The resulting distribution of numerical values of ε_{eff} "in and around the protein" for the field produced by the charge ⊕ and the induced polarized charges looks as shown in Fig. 6.2A if ⊕ is in water near the protein surface, and as in Fig. 6.2B if it is deep in the protein. Let me remind you that ε_{eff} is the effective value of permittivity for the point $\boldsymbol{r}$ to be used in the formula $\varphi(\boldsymbol{r}) = q_1/[\varepsilon_{\text{eff}}|\boldsymbol{r}\text{-}\boldsymbol{r}_1|]$ to calculate the potential of the charge ⊕ located at $\boldsymbol{r}_1$.

It is useful to have in mind a simple expression describing interaction of two charges in a nonuniform media:

$$U \approx \frac{q_1}{\langle\varepsilon\rangle_1} \times \frac{\langle\varepsilon\rangle_{12}}{r} \times \frac{q_2}{\langle\varepsilon\rangle_2} \tag{6.4}$$

where $\langle\varepsilon\rangle_1$ is the average permittivity around charge q_1 (so that $\dfrac{q_1}{\langle\varepsilon\rangle_1}$ is this charge partly screened by polarization of the medium), $\langle\varepsilon\rangle_2$ is that for $q_2$vicinity, and $\langle\varepsilon\rangle_{12}$ is the average permittivity in-between and around charges q_1, q_2. As seen, in a uniform medium (when $\langle\varepsilon\rangle_1 = \langle\varepsilon\rangle_2 = \langle\varepsilon\rangle_{12} = \varepsilon$), this expression is reduced to a conventional form given by Eq. (6.2).

It is worth considering one more consequence of the interface. This is the effect of the charge on itself: a charge located outside the protein is *repelled* from the protein surface, while a charge inside the protein is *strongly attracted* to the protein surface, that is, actually, to water. In both cases, the medium of higher permittivity attracts the charge, and that with a lower permittivity repels it. The values of these effects can be estimated (Finkel'shtein, 1977).

Let us consider some numerical examples. One can show that the energy of an ion (modelled as a conducting sphere with a charge q equal to the proton charge and radius $R = 1.5$ Å) is $q^2/[2\varepsilon_1 R] \approx 1.4$ kcal mol^{-1} when it is in water

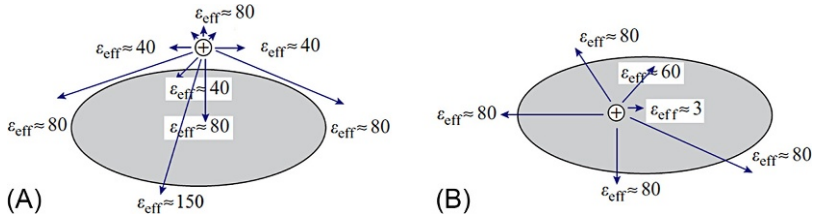


FIG. 6.2 Typical effective permittivity values ε_{eff} in various points for a potential produced by the charge ⊕ located near the protein surface (A) and inside the protein (B). In both cases, the protein (with permittivity 3) is surrounded with water (with $\varepsilon \approx 80$).

with permittivity $\varepsilon_1 \approx 80$ and $q^2/[(\varepsilon_1+\varepsilon_2)R] \approx 2.7$ kcal mol^{-1} when it is half-immersed in the protein with permittivity $\varepsilon_2 \approx 3$ (see Problem 6.5). Thus, the half-immersion-caused increase in energy is ≈ 1.3 kcal mol^{-1}. But when the same charge is placed between two proteins (or just deeply immersed in a protein), its energy is $q^2/[2\varepsilon_2 R] \approx 36.9$ kcal mol^{-1}. Thus, the full-immersion-caused increase in energy is ≈ 35.5 kcal mol^{-1}, which is nearly 30-fold energy of half-immersion. This illustrates a strong nonadditivity of electrostatic interactions in a nonuniform medium.

Now let us focus on the effects connected with the molecular structure of the medium. Polarization of the molecules determines the value of permittivity ε. If the medium consists of nonpolar molecules, an electric field only shifts electrons in the molecules, which is rather difficult; therefore, the polarization is small, and ε is low. If the medium consists of polar molecules (eg, of waters), an electric field turns these molecules, which is easier, and, hence, ε of such medium is high.

In both cases, with electrons shifted or molecules turned, polarization of the medium partially screens the immersed charges ($\oplus$ and $\ominus$, see Fig. 6.3A) and thereby diminishes the electric field in the medium compared with what it would have been in a vacuum.

It would be only natural to expect that the polarized molecules must strongly affect the interaction of charges at short distances, since the classical equations, such as (6.1)–(6.3), are valid, strictly speaking, only when the charges are separated by many medium molecules. And if the charges are 3–4 Å apart (as often happens in proteins), no other atom can get between them to change their interaction.

In the case of such close contact, the permittivity might be believed to approach 1, even in the aqueous environment. This viewpoint, or rather this misapprehension, may still be encountered in the literature.

However, strange as it may seem, the medium's particulate nature makes no drastic changes in the "macroscopic" (ie, derived for large distances between the charges) permittivity, even if the distance is as short as 3 Å. In other words, even at the smallest distances (when $\oplus$ and $\ominus$ are in a close contact), the value of water permittivity is much closer to 80 or 40 than to 1 or 3.

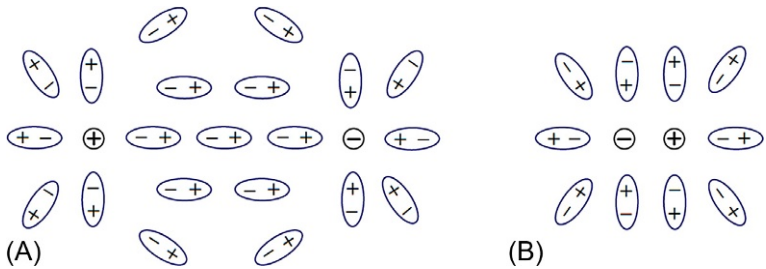
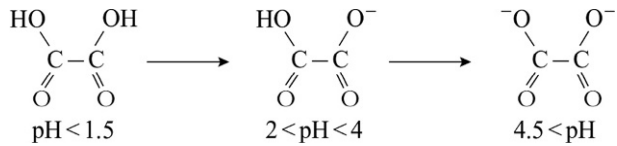


FIG. 6.3 Schematic drawing of orientation of water molecules surrounding charges $\oplus$ and $\ominus$ which are not in (A) or are in (B) a close contact.

This follows from the fact that salt easily dissolves in water, which is possible only with a weak attraction between counter-ions, even at very short distances of about 3 Å (Finkel'shtein, 1977).

Indeed, the distance between Na^+ and Cl^- ions is as short as 3 Å in a direct van der Waals contact. Then their free energy of attraction would amount to -1.5 kcal mol^{-1} at $\varepsilon = 80$, -3 kcal mol^{-1} at $\varepsilon = 40$, and -6 kcal mol^{-1} at $\varepsilon = 20$. The latter (6 kcal mol^{-1}) exceeds the energy of a hydrogen bond. Such an energy would make the counter-ions stick together more tightly than water molecules do, and then the concentration of a saturated salt solution would be about 10^{-4} mol L^{-1}, like the concentration of saturated water vapor. But this obviously cannot be true: it is no problem to dissolve one mole of NaCl (58 g) in 1 L of water (this will be an ordinary, perhaps a bit too salty brine). Consequently, water permittivity is considerably higher than 20, even at a distance of about 3 Å.

The value of ε at the closest distances inside the molecule can be estimated more precisely using the first and second constants of dissociation of dihydric acids in water. For example, dissociation of oxalic acid occurs as follows:



The second dissociation is shifted from the first one by approximately 2.5 pH units, that is, it occurs when the H^+ concentration is $10^{2.5} = e^{2.3\times2.5}$ times lower. This shows that the free energy of interaction of the first charge with the second one is $2.5 \times 2.3RT \approx 3.5$ kcal mol^{-1} when the distance between the charges is about 2.5–3 Å. This value of interaction energy corresponds to $\varepsilon \approx 40$ at a distance of 2.5–3 Å. A similar result ($\varepsilon \approx 30$ at a distance of about 2–2.5 Å) was obtained for dissociation of carbonic acid, $H_2CO_3 \rightarrow HCO_3^- \rightarrow CO_3^{2-}$, and of other dibasic acids and bases (Birshtein et al., 1964). It should be noted that a "correct" hydrogen bond energy is obtained at $\varepsilon \approx 20$ at a distance of about 2 Å (and, of course, ε must approach 1 at a distance of ≈ 1 Å—otherwise energy of all covalent interactions would change drastically, which is not observed).

Hence, even a salt bridge between oppositely charged side-chains on the protein surface must "cost" 2–3 kcal mol^{-1} only. In the interior of the protein, its "price" is higher, but immersion of charged side-chains into the protein would have been still more expensive, so it is no wonder that such bonds are rather rare in native proteins.

Thus, we conclude that the particulate nature makes no drastic changes in the "macroscopic" (derived for large distances between the charges) permittivity of water even at a distance of about 3 Å, which is too small for any other molecule to get between the interacting charges. The reason is suggested by the already considered case of "a charge and a dipole," which suggests that a dipole located between charges increases rather than decreases their interaction, and thus this central dipole does not increase the value of permittivity. The permittivity is increased not by the central but by other dipoles. Thus, we can conclude that the charges are quite well shielded by solute molecules coming from other sides and from the flanks (Fig. 6.3B): these molecules become polarized (in the case of water, they simply turn), so that their "+"s shift towards the charge $\ominus$, and their "−"s towards the charge $\oplus$.

Here we see again (see also Fig. 6.2A and B) that the electrostatic interaction between the charges occurs mainly via the medium of a higher permittivity and nearly ignores the medium of weak polarization.

All our previous discussions referred to "micromolecular" systems. But are the conclusions drawn valid for proteins (where the particulate effects are coupled with the huge difference in permittivity between the water and the protein)?

Experiments reported by the research team headed by A. Fersht, the founder of protein engineering, show that the above estimates are valid for proteins (Russel and Fersht, 1987; Fersht, 1999).

The basis of these experiments is as follows. There are proteins (enzymes) that exhibit a particularly high activity at a certain value of pH (Nelson and Cox, 2012); they are said to have a pH-optimum. This pH-optimum can be shifted (Fig. 6.4) using a charged residue introduced to the protein by mutating its gene, and the electric field induced at the active site by the charge of the mutated residue can be estimated by the shift of the pH-optimum.

The pH-optimum is caused by the fact that, to keep the enzyme functioning, a group at its active site must have a certain charge that, in turn, depends on the concentration of hydrogen ions in the medium. The H^+ concentration (ie, $[H^+]$ mol L^{-1}) is equal to 10^{-pH} by definition, and the OH^- concentration in water is about 10^{-14+pH}.

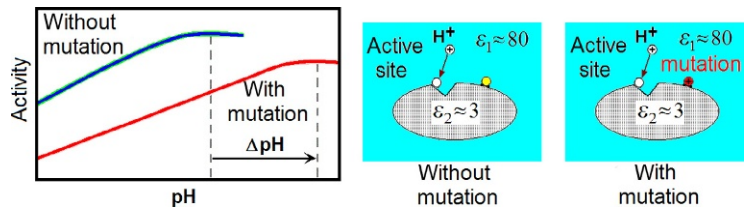


FIG. 6.4 A scheme of experiment on the mutation-caused shift of pH-optimum (after Russel and Fersht, 1987, with simplifications).

Let the active site (AS) accept the ion H^+: $AS + H^+ = ASH^+$. Then, according to the active mass law, the ratio between these two forms of the active site (with and without H^+) is:

$$\begin{aligned}\frac{[ASH^+]}{[AS]} &= \exp\left(-\frac{\Delta F_{ASH^+}}{RT}\right) \times [H^+] = \exp\left(-\frac{\Delta F_{ASH^+}}{RT}\right) \times 10^{-pH} \\ &= \exp\left\{-\left(\frac{\Delta F_{ASH^+}}{RT} + 2.3 \times pH\right)\right\}\end{aligned} \tag{6.5}$$

where ΔF_{ASH^+} is the free energy of H^+ binding (at $[H^+] = 1$ mol L^{-1}) to the active site, and the symbol $[\cdot]$ denotes concentration.

If the mutation-introduced charge induces a potential φ at the protein active site, then ΔF_{ASH^+} changes as $\Delta F_{ASH^+}|_{\text{with mutation}} = \Delta F_{ASH^+}|_{\text{without mutation}} + \varphi \times e$, where e is the charge of H^+. Since at the pH-optimum the magnitude of $\frac{[ASH^+]}{[AS]}$ (and the magnitude of $\frac{\Delta F_{ASH^+}}{RT} + 2.3 \times pH$) must remain the same both with and without the mutation, then

$$\begin{aligned}&\frac{\Delta F_{ASH^+}|_{\text{without mutation}}}{RT} + 2.3 \times pH|_{\text{opt. without mutation}} \\ &\frac{\Delta F_{ASH^+}|_{\text{with mutation}}}{RT} + 2.3 \times pH|_{\text{opt. with mutation}}\end{aligned} \tag{6.6}$$

That is,

$$\begin{aligned}&\varphi \times e = \Delta F_{ASH^+}|_{\text{with mutation}} - \Delta F_{ASH^+}|_{\text{without mutation}} \\ &= 2.3RT \times \left(pH|_{\text{opt. without mutation}} - pH|_{\text{opt. with mutation}}\right) = 2.3RT \times (-\Delta pH)\end{aligned} \tag{6.7}$$

Thus, having learned the shift of the pH-optimum, ΔpH, we can estimate the potential induced at the active site by the mutated protein residue. Then, using the known three-dimensional structure of the protein, and, hence, the distance r from the mutated residue to the active site, we can estimate the effective permittivity ε_{eff} (a term in the equation $\varphi = q/(\varepsilon_{\text{eff}} r)$) for the interaction between the mutation-introduced charge q and the active site.

In Fersht's experiments, the mutations were performed at the surface of the protein in order not to damage its structure (as we have already learned, the energy of a charge deeply immersed in the protein is high and can literally explode the protein globule).

The experimental result reported: the effective permittivity ε_{eff} ranges from about 40 to 100, the former being typical of mutations at short distances from the active site, and the latter for remote (and shaded by the protein body) ones. The fact that ε_{eff} can reach a value of 100 appeared to be not a little surprising to those believing that ε_{eff} must lie between 3 (as inside the protein) and, at the

most, 80 (as in water). However, for us, these values are not surprising, since they are in good agreement with what follows from Fig. 6.2A.

A brief digression on protein engineering. Its major advantage is that by changing a codon in a protein gene, we can perform a mutation at an exact site of the protein globule, since the gene in question, the amino acid sequence of the protein and the protein 3D structure are known. Besides, the mutation effect on the structure can also be monitored by X-ray analysis or NMR. Thus, the entire work is performed with one's eyes open.

In the experiments that we discussed earlier, the protein served as a microscopic (or rather, nanoscopic) electrometer. And protein engineering allows us to use such instruments as well as jumping from the physical theory to genetic manipulations, which is great fun!

Here, I would like to make some additions concerning electrostatic interactions.

First: So far, I have discussed only the interaction of separate charges. However, electrostatics also covers the interactions of dipoles (eg, the dipoles $H^{(+)}$–$O^{(-)}$ and $H^{(+)}$–$N^{(-)}$ involved in hydrogen bonding) as well as quadruples; the latter are present, for example, in aromatic rings (Fig. 6.5).

The reason for my considering interactions of ions is simply the strength of these interactions: even in immediate contact they are a few times stronger than interactions between dipoles (also, their decrease with increasing distance is slower), while interactions of dipoles are stronger than those between quadrupoles.

Second: With free charges (eg, salt) available in the water, the electrostatic interactions diminish with distance r as $(1/r) \times \exp(-r/D)$ rather than as $1/r$ [Pauling, 1970]. Here D, the Debye-Hückel radius, corresponds to the typical size of the counter-ion cloud around the charge. The value of D is independent of the charge itself but depends on the charges of the salt ions, on their concentration in the medium, on its permittivity and on temperature. In water, at room temperature:

$$D \approx \frac{3}{I^{1/2}} \text{Å} \tag{6.8}$$

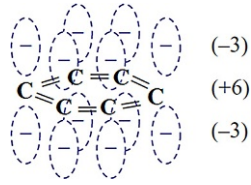


FIG. 6.5 The electric quadruple of an aromatic ring: the layer of "halves" of six p-electrons (charge −3)—the layer of cores (charge +6)—the layer of the last "halves" of p-electrons (charge −3).

where

$$I = \frac{1}{2}\sum_i c_i z_i^2 \tag{6.9}$$

is the *ionic strength* of the solution given in moles per liter. In Eq. (6.9), the sum is taken over all kinds of ions present in the solution, z_i is the charge (in proton charge units), c_i is the concentration (in moles/L) of ion i. Under ordinary physiological conditions (room temperature, $c_i \approx 0.1$–0.15 mol L^{-1} for ions with the charges +1 and −1, while the concentration of ions with the charge +2 is much lower), $I \approx 0.1$–0.15 mol L^{-1}, and then $D \approx 8$ Å. However, some microorganisms live at $I \approx 1$ mol L^{-1} and more; then the persisting electrostatic interactions of the charged groups of the protein, though much weakened, are only those corresponding to "salt bridges", that is, to the immediate contact of the charges.

In general, with an ionic atmosphere present in the solution, the energy of interaction of the two charges is

$$U = \frac{q_1 q_2}{\varepsilon_{\text{eff}} r} \times \exp\left(-\frac{r}{D}\right) \tag{6.10}$$

Third: As seen from above, the electrostatic interaction is a striking example of the *non*-pairwise interaction of particles (unlike, for example, van der Waals interaction). It depends not only on the distance r between the charges q_1 and q_2 but also on the medium properties (those change both ε and D), and specifically, on the distance between the charges and other bodies and on the shape of these bodies (these affect ε_{eff}), as well as on the concentration of free ions with the charges +1 and −1 in the solution (which affects D). Nevertheless, the maths form for the resulting interaction is rather simple.

One more example of this kind: The energy of interaction of a charge q situated in a salt-free medium with permittivity ε_1 with a small noncharged body of volume V and permittivity ε_2 can be estimated (Finkel'shtein, 1977) as:

$$U \approx \frac{q^2 V}{8\pi r^4} \times \frac{\varepsilon_1 - \varepsilon_2}{\varepsilon_1\left(\frac{2\varepsilon_1}{3} + \frac{\varepsilon_2}{3}\right)} \tag{6.11}$$

And one more addition: So far, I have used the term "the energy of electrostatic interactions." This was done for the sake of simplicity; as I mentioned at the beginning of this lecture, the strict term would be "free energy." This is because we focused only on the attraction and repulsion of charges, without preventing the *heat exchange* with the environment. And with heat exchange allowed, what we dealt with is the free energy by definition.

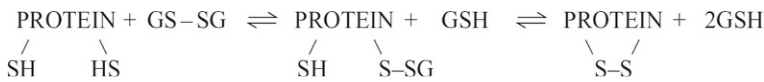
Moreover, the temperature dependence of electrostatic effects in aqueous environment may be used to show the predominance of the entropic constituent over the energetic (enthalpic) one; the latter, by the way, is close to zero. This

follows from the fact that water permittivity decreases from 88 to 55 (ie, the electrostatic interactions increase by 40%), with increasing absolute temperature T from 273 to 373 K (ie, by 35%). This means that the electrostatic interactions increase approximately proportionally to the absolute temperature. And the interaction increasing proportionally to the absolute temperature implies an exclusively entropic constituent (take a look at Problem 6.7). Hence, in water, the entire electrostatic effect is caused *not* by energy, but by the ordering of water molecules around the charges and by its variation with variation distance between the charges.

Hence, strange as it may seem, electrostatics in water originates *not* from energy but from entropy, just like hydrophobic interactions or hydrogen bonding in an aqueous environment.

In concluding this section on "Elementary interactions in and around proteins," I would also like to mention disulfide and coordinate bonds. Although not as abundant in proteins as hydrogen bonds, these may often be of great importance (Schulz and Schirmer, 1979, 2013; Creighton, 1993).

Disulfide (S–S) bonds are formed by cysteine (Cys) amino acid residues (the Cys side-chain is $-C^{\beta}H_2-SH$). No direct oxidation of cysteines accompanied by hydrogen release (according to the scheme $-CH_2-SH + HS-CH_2- \rightarrow -CH_2-S-S-CH_2- + H_2$) occurs in proteins because at room temperature this process is too slow. However, in proteins, S–S-bonding can be rapid when assisted by thiol-disulfide exchange. In the cell, the exchange is thought to involve *glutathione* that has both the monomeric thiol (GSH) and dimeric disulfide (GSSG) form, and to follow the scheme



Both breakdown and formation of S–S-bonds can occur spontaneously *in vitro*, but in cells, these are catalysed (ie, accelerated but not directed) by a special enzyme, disulfide isomerase.

Rotation about the S–S bond is rather hindered by a high torsional potential that only allows torsional angles of about +90 degree and −90 degree (Creighton, 1993).

S–S-bonding is reversible, since the energetic equilibrium of this reaction (thiol-disulfide exchange) is close to zero (there were two covalent S–H bonds and one S–S bond, and now there are as many; this resembles the energy balance of hydrogen bonding in proteins, does it not?). Moreover, the available (rather high) concentration of GSH in the cell shifts the equilibrium towards breaking the bonds that might be produced by an "occasional" cysteine approach.

Therefore, the only S–S bonds that can be formed and persist are between cysteines brought close to one another by other interactions (Creighton, 1993; Sevier and Kaiser, 2002).

S–S bonds are of particular importance for proteins that have to reside and function out of the cell. On the one hand, the absence of disulfide isomerase and glutathione is favorable for bonds that have been already formed (either in the cell or when leaving it) as they become "frozen" and run no risk of breaking or rearranging. On the other hand, the external conditions may be different, and the extended margin of safety provided by the stable, "frozen" S–S bonds would not be out of place for the protein. That is why S–S bonds are typical of secreted proteins rather than cellular ones (Creighton, 1993; Sevier and Kaiser, 2002). Usually, in secreted proteins, all available cysteines (less one, if their number is odd) participate in S–S bonds.

Coordinate bonds are formed by N, O, and S atoms of the protein (as well as by O atoms of water) to di- and trivalent ions of Fe, Zn, Co, Ca, Mg, and other metals.

The ions of these metals have vacant orbitals lying low (as to energy), only slightly above their filled electron orbitals. Each of the vacant orbitals is capable of bonding an electron pair. And O, N, and S atoms (electron donors) have electron pairs that can occupy the vacant orbitals of the ions. The resultant bonds are identical to ordinary chemical bonds, except that ordinary bonds comprise electrons from both parent atoms, while coordinate bonds are formed by electrons coming from one bonded atom only (Pauling, 1970).

During coordinate bonding, the metal ion binds to several donors of electrons. Then a small (radius ~0.7 Å) di- or trivalent ion is surrounded by large-atom donors (radius ~1.5 Å). Mostly, there are six of these coordinating donor atoms located at the apices of a regular octahedron (Pauling, 1970) (Fig. 6.6).

Since the ion can be bonded to both electron donors of the protein and to oxygens of water, it passes (despite the high energy of each bond) from water

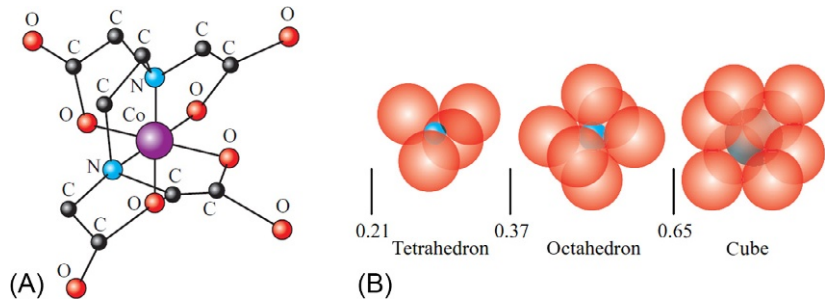


FIG. 6.6 (A) The structure of the octahedral complex between Co^{3+}-atom and EDTA. (B) Typical coordination of the central ion at various ratios between its radius and the radii of the surrounding electron donors. *(Adapted from Pauling, L., 1970. General Chemistry. W.H. Freeman & Co., New York (Chapter 19).)*

to the protein and back without dramatic gains or losses in energy. What is of greater importance is, if the positions of the donor atoms in the protein are "proper" for coordinate bonding, the ion can release the water molecules of its previous water environment and bind to the protein; then a strong bond occurs owing to the gained entropy of motion of the released water molecules (very much like the energy and entropy balance of hydrogen bonding, is it not?). On average, each coordinate bond costs several kilocalories per mole, that is, it is a little more expensive than a hydrogen bond in a water environment.

Such coordinate bonds formed by several atoms of one molecule are called *chelate* (claw-shaped). The role of these bonds in proteins, specifically at their active sites, will be considered later on. Also, we will see later, that chelate complexes coating ions completely can be members of the hydrophobic protein core. At the moment I would like to draw your attention again to Fig. 6.6 presenting the widely used reagent EDTA (ethylenediaminetetraacetic acid) that participates in a chelate bond to the metal. For EDTA, this bond is particularly strong because the negatively charged COO^- groups of EDTA are bound to the positively charged metal ion.

REFERENCES

Birshtein, T.M., Ptitsyn, O.B., Sokolova, E.A., 1964. Theory of polyelectrolytes. V. Interactions of near groups in stereoregular polyelectrolytes. Vysokomol. Soedin. (in Russian) 6, 158–164.

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W. H. Freeman & Co., New York (Chapters 1, 2, 7).

Donchev, A.G., Galkin, N.G., Illarionov, A.A., Khoruzhii, O.V., Olevanov, M.A., Ozrin, V.D., Pereyaslavets, L.B., Tarasov, V.I., 2008. Assessment of performance of the general purpose polarizable force field QMPFF3 in condensed phase. J. Comput. Chem. 29, 1242–1249.

Fersht, A., 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman & Co., New York (Chapter 5).

Finkel'shtein, A.V., 1977. Electrostatic interactions of charged groups in aqueous medium and their influence on the formation of polypeptide secondary structures. Mol. Biol. (Moscow) 11, 811–819.

Halgren, T.A., 1995. Merck molecular force field. I. Basis, form, parameterization and performance of MMFF94. J. Comput. Chem. 17, 490–519.

Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J., 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J. Am. Chem. Soc. 118, 11225–11236.

Landau, L.D., Lifshitz, E.M., Pitaevskii, L.P., 1984. Electrodynamics of Continuous Media, second ed. A Course of Theoretical Physics, Volume 8. Elsevier Butterworth-Heinemann, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo. §§ 6–8, 11.

Levitt, M., Hirshberg, M., Sharon, R., Daggett, V., 1995. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. Comput. Phys. Commun. 91, 215–231.

MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K.,

Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher III, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B. 102, 3586–3616.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co., New York (Chapter 2, 6).

Pauling, L., 1970. General Chemistry. W.H. Freeman & Co., New York (Chapter 19).

Pereyaslavets, L.B., Finkelstein, A.V., 2012. Development and testing of PFFSol1.1, a new polarizable atomic force field for calculation of molecular interactions in implicit water environment. J. Phys. Chem. B. 116, 4646–4654.

Russel, A.J., Fersht, A.R., 1987. Rational modification of enzyme catalysis by engineering surface charge. Nature 328, 496–500.

Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapters 3, 4).

Sevier, C.S., Kaiser, C.A., 2002. Formation and transfer of disulphide bonds in living cells. Nat. Rev. Mol. Cell Biol. 3, 836–847.

Wang, J., Wolf, R.M., Caldwell, J.W., et al., 2004. Development and testing of a general amber force field. J. Comput. Chem. 25, 1157–1174.

Part III

Secondary Structures of Polypeptide Chains

This page intentionally left blank

Lecture 7

Having dealt with *elementary* interactions, in this lecture we consider the secondary structure of proteins and polypeptides. It is well studied (see Schulz and Schirmer, 1979, 2013; Cantor and Schimmel, 1980; Branden and Tooze, 1999; Creighton, 1993).

First of all, we will discuss regular secondary structures, that is, α-helices and β-structures (Pauling and Corey, 1951a,b; Pauling et al., 1951). These secondary structures are distinguished by regular arrangements of the main chain with side chains of a variety of conformations. The tertiary structure of a protein is determined by the arrangement of these structures in the globule (Fig. 7.1).

We shall consider helices first. They can be right-handed or left-handed (Fig. 7.2) and have different periods and pitches. Right-handed (R) helices come closer to the viewer as they move counterclockwise (which corresponds to positive angle counting in trigonometry), while left-handed (L) helices approach the viewer as they move clockwise.

In the polypeptide chain, major helices are stabilized by hydrogen bonds.

The bonds are formed between C═O and H–N groups of the polypeptide backbone, the latter being closer to the C-terminus of the chain.

In principle, the following H-bonded helices can exist (Fig. 7.3): 2_7, 3_{10}, 4_{13} (usually called α), 5_{16} (called π), etc. The 2_7-helix derives its name from the second residue participating in the H-bond (see Fig. 7.3) and seven atoms in the ring ($O \rightarrow H–N–C'–C^{\alpha}–N–C'$) closed by this bond. Other helices (3_{10}, 4_{13}, etc.) are named accordingly.

Which of these helical structures are most abundant in proteins? α-Helices are. Why? This question is answered by the Ramachandran map for a typical amino acid residue, alanine (Fig. 7.4), where I marked the conformations that, being repeated periodically, cause formation of the H-bonds shown in Fig. 7.3 (Ramachandran and Sasisekharan, 1968; Schulz and Schirmer, 1979, 2013).

As seen, only the α_R-helix (right-handed α-helix) is deep inside the region allowed for alanine (and for all other "normal," ie, L amino acid residues). Other helices are either at the very edge of this allowed region (eg, the left-handed α_L-helix or the right-handed 3_{10}-helix), which gives rise to conformational strains, or in the region allowed for glycine only.

Therefore, it may be expected that the right-handed α-helix is the most stable and hence (*what's good for General Motors is good for America*) most abundant in proteins—and this is really so. In the right-handed α-helix (Fig. 7.5), the arrangement of all atoms is optimal, ie, tight but not strained. Therefore, it is no wonder that in proteins α-helices are numerous, and in fibrous proteins, they are extremely extended and incorporate hundreds of residues.

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00007-7

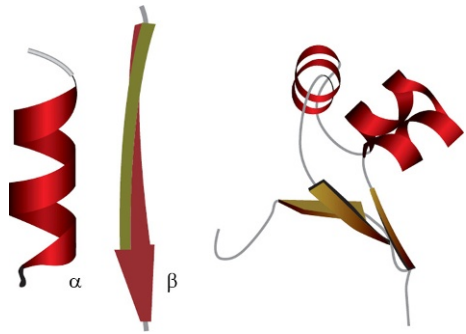


FIG. 7.1 The secondary structures of a polypeptide chain (α-helix and a strand of β-sheet) and the tertiary structure of a protein globule (*on the right*). Usually, taken together, α- and β-structures make up about a half of the chain in a globular protein.

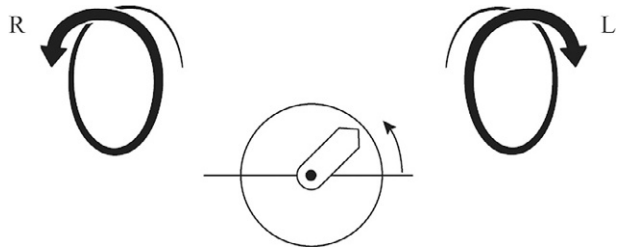


FIG. 7.2 Right-handed (R) and left-handed (L) helices. The spiral naturally must be viewed along its axis, and, with equal success, from either of its ends. The bottom part of the figure shows positive angle counting in trigonometry: the *arrow* that is "close" to the viewer moves *counterclockwise*.

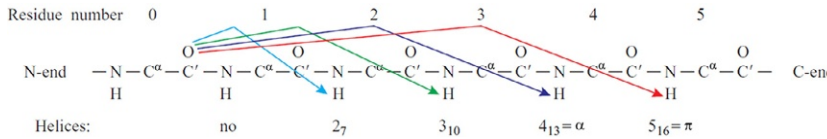


FIG. 7.3 Hydrogen bonds (shown with *arrows*) typical of different helices. The chain residues are numbered from the N- to the C-end of the chain.

Left-handed α-helices are not (or hardly ever) observed in proteins. This is also true for 2_7-helices that not only lie at the very edge of the allowed region but also have a large angle between their N–H and O=C groups, that is energetically disadvantageous for hydrogen bonding. π-Helices are absent from proteins too. They also occur at the very edge of the allowed region and their turns are far too wide, which results in an energetically unfavorable axial "hole." In contrast, 3_{10}-helices (mainly right-handed; left-handed ones are good for glycines only) are present in proteins, although only as short (three to four residues) and distorted fragments: the 3_{10}-helix is too tight and gives rise to steric strains; its conformation lies close to the edge of the allowed region.

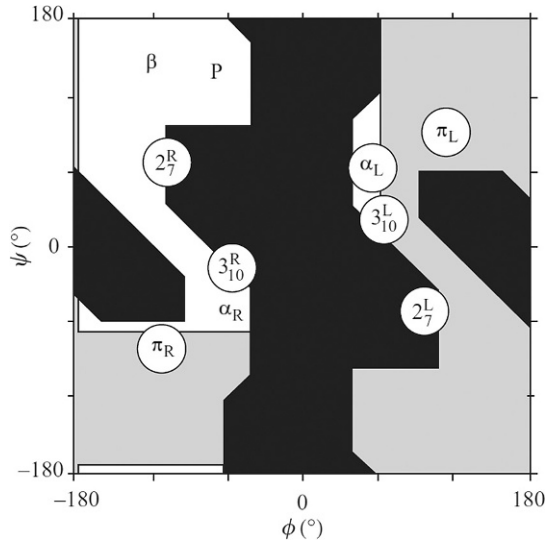


FIG. 7.4 The conformations of various regular secondary structures against the map of allowed and disallowed conformations of amino acid residues. 2_7^R, 2_7^L: the right-handed and left-handed 2_7-helix; 3_{10}^R, 3_{10}^R: the right-handed and left-handed 3_{10}-helix; α_R, α_L: the right-handed and left-handed α-helix; π_R, π_L: the right-handed and left-handed π-helix. β, the β-structure (for details, see Fig. 7.8B). P, the poly(Pro)II-helix. □, conformations allowed for alanine (Ala); ▒, regions allowed for glycine only, but not for alanine and other residues; ■, regions disallowed for all residues; ϕ and ψ, dihedral angles of rotation in the main chain.

Pay attention to the feature clearly seen in Fig. 7.5A: the helical N-terminus is occupied by "free" H atoms of N–H groups uninvolved in intra-helical H-bonds, while the C-terminus is occupied by H-bond-free O atoms of C=O groups. Since the electron cloud of the H atom is partially pulled off by the electronegative N atom of NH group, and the electronegative O atom attracts the electron of the C′ atom of C′O group, the N-terminus assumes a positive and the C-terminus a negative partial charge (Ptitsyn and Finkel'shtein, 1970). That is, the helix is a long dipole where the N-terminal partial "+" charge (coming from three "H-bond-free" H-atoms) amounts to about half of the proton charge, while the C-terminal "−" charge amounts to about half of the electron charge.

Now let us consider the regular main-chain structures lacking hydrogen bonds inside each of them but periodically H-bonded with one another.

The extended (all angles in the main chain are nearly *trans*), slightly twisted chains ("β-strands") form the sheet of the β-structure. A β-sheet can be (Fig. 7.6) parallel, (β ↑↑) antiparallel (β ↓↑), and mixed (comprising β ↑↑ and β ↓↑). The β-structure is stabilized by H-bonds (shown in light-blue lines in Fig. 7.6). Since the surface of -structure sheets is pleated (Fig. 7.7), this structure is also called the "pleated β-structure."

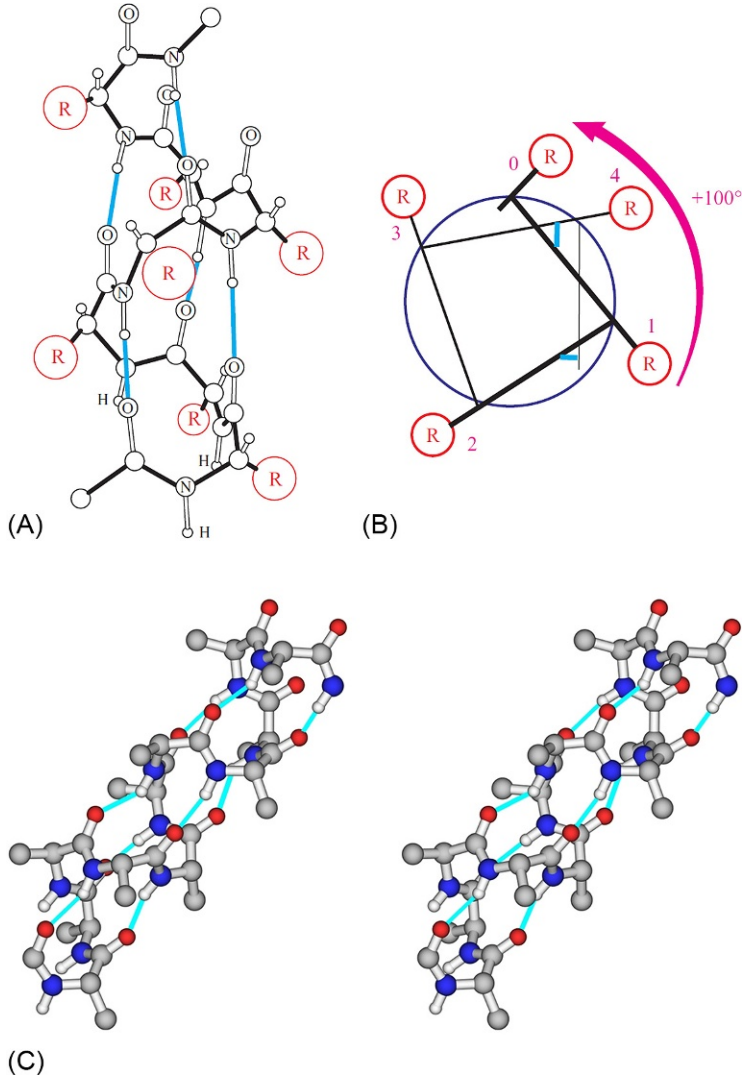


FIG. 7.5 The right-handed α-helix. Hydrogen bonds in the main chain are shown as *light-blue lines*. (A) Atomic structure; R = side-chains. (B) Axial view of one turn of this α-helix. The *arrow* shows the turn of the helix (per residue) when it approaches the viewer (the closer to the viewer, the smaller the chain residue number). The circle depicts the cylindrical surface enveloping the C^{α} atoms of the helix. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5). (C) Stereo drawing (see Appendix E) of an α-helix. In side chains, only C^{β} atoms are shown.)*

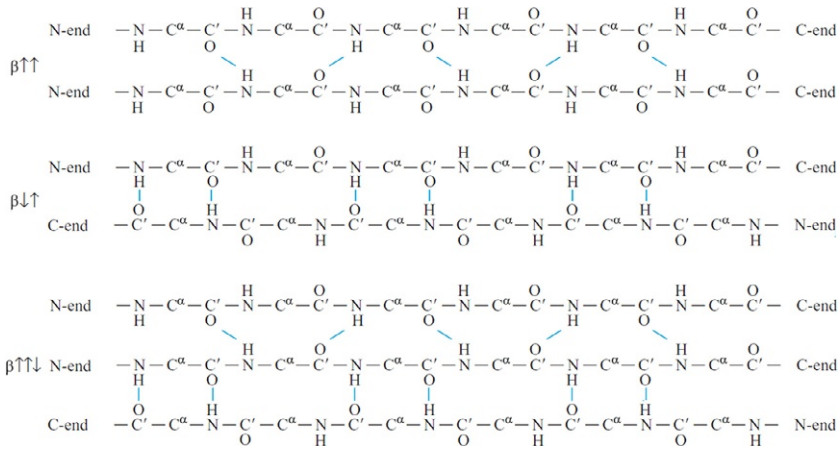


FIG. 7.6 Chain pathway and location of hydrogen bonds in the parallel (β ↑↑), antiparallel (β ↓↑), and mixed (β ↑↑↓) β-structures. As shown, in each β-strand, the H-bonds (*light-blue*) of one residue are directed oppositely to those of its neighbor in the chain. Periodicity of any β-structure equals two.

As a whole, the β-sheet is usually somewhat twisted (Figs. 7.7A,B, and 7.8A) because each separate β-strand is twisted by itself (Fig. 7.8B), thereby slightly altering the direction of H-bonds along the strand and thus favoring a certain angle between the H-bonded strands. In its turn, as explained by Chothia (1973), this twist of a strand results from shifting the energetically advantageous conformation of all side-chain-possessing residues towards the center of the sterically allowed region (Fig. 7.8C). The twist of an individual β-strand is left-handed (for "normal," ie, L amino acids; for D, it would be opposite); as seen from Fig. 7.8B, the strand's side chains turn clockwise (by about −165° per residue) as the strand comes closer to the viewer.

Because the strands twist, H-bonds turn as well (by about −165° per residue, ie, by −330° = +30° per residue pair, a regular periodic element of the β-structure). As a result, the angle between the neighboring β-strands (viewed from the edge of the sheet, Fig. 7.8A) usually amounts to about −25° (as always, "−" means a clockwise turn of the nearby β-strand about the remote one). Thus, the β-sheet has a left-handed twist if viewed from its edge (and *right*-handed if viewed along the β-strands, as is usually done).

There are also helices without any hydrogen bonds. Their tight (and hence, energetically advantageous) arrangement is stabilized by van der Waals interactions only. This is exemplified by a polyproline helix consisting of three chains; each chain forms a rather extended left-handed helix. Winding these three chains together forms a *triple right-handed superhelix*. Of two possible types of polyproline helix, that of interest to us is poly(Pro)II (Traub and Piez, 1971), since a helix of this kind is realized in collagen. In this helix, the Pro-peptide groups are in the usual (*trans*)conformation. Let us postpone a detailed consideration of the collagen helix until another time and at the

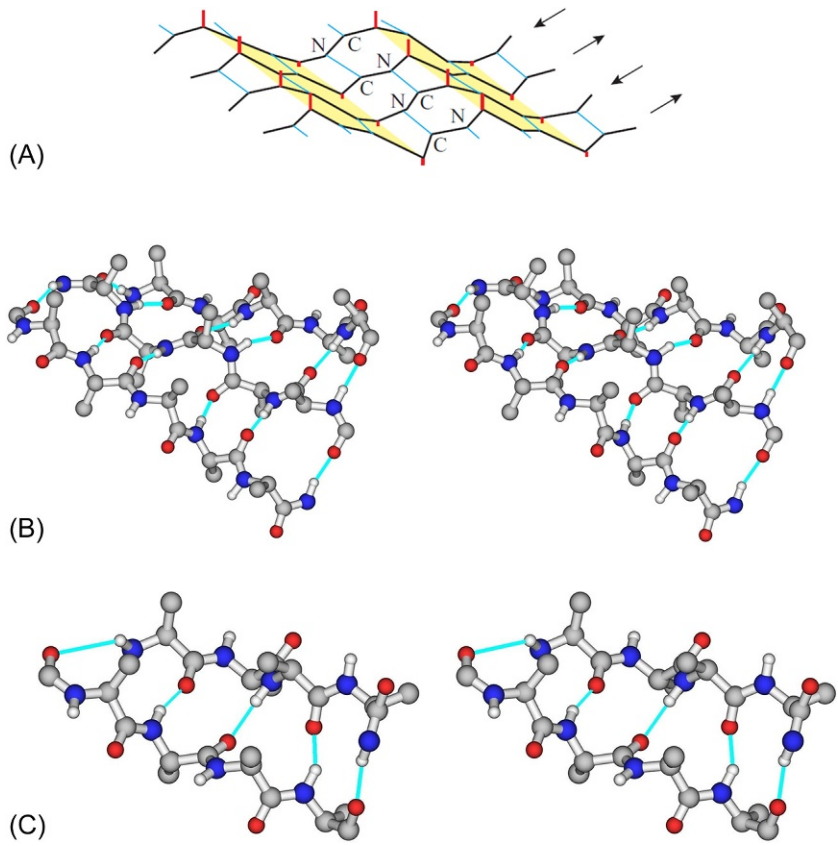


FIG. 7.7 (A) An idealized (nontwisted) antiparallel β-sheet. The β-sheet surface is pleated. The side chains (shown as short *red rods* here) are at the pleats and directed accordingly; ie, the upward and downward side chains alternate along the β-strand. The H-bonds are shown in *light-blue*. Adapted from Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5). Below: stereo drawings of β-sheets taken from real protein structures. These β-sheets are twisted a little: (B) an antiparallel β-sheet of three β-strands; (C) a parallel β-sheet of two β-strands.

moment, restrict ourselves to its overall view (Fig. 7.9) and mark in Fig. 7.4 the region P corresponding to its conformation: one can see that it is close to the β-structural conformation.

The features of major regular secondary structures of protein chains are summed up in Table 7.1.

Apart from the regular standard secondary structures, in polypeptide chains there are irregular ones, ie, standard structures that do not form long periodic systems.

The most important are the so-called β-turns (Venkatachalam, 1968) (or β-bends; "β" in the name shows that they often bridge the neighboring β-strands

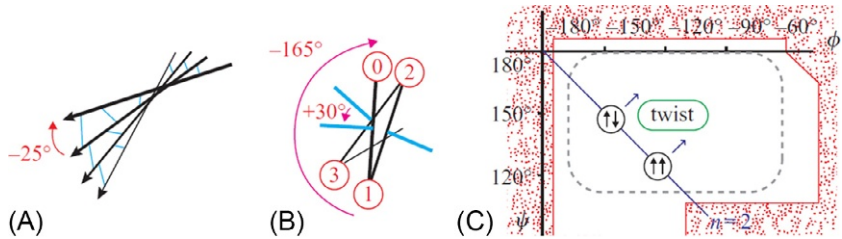


FIG. 7.8 (A) The twist of the β-sheet. β-strands are shown as *arrows* and hydrogen bonds between them as *light-blue lines*. (B) Axial view of one turn of the β-strand. Side chains are shown as *circles*; their numbers increase with increasing distance from the viewer. *Blue lines* indicate the direction of C═O groups involved in H-bonding in the sheet. The *large arrow* shows the turn of the β-strand as it comes closer to the viewer by one residue, and the *small arrow* shows the turn of similarly directed H-bonds when the β-strand comes closer to the viewer by two residues. (C) Conformation of the ideal (nontwisted) parallel (↑↑) and antiparallel (↑↓) β-structure for poly(Gly), and the averaged conformation of a real twisted β-structure (composed of L amino acids). The *dashed line* encircles the energy minimum for a separate Ala; the allowed region for its conformations is contoured by the *red line*. The diagonal of the *φψ*-map corresponds to the flat regular structure with two residues per turn. Left-handed (L) helices are above the diagonal, and right-handed (R) ones are below it (Ramachandran and Sasisekharan, 1968). *(Parts (A) and (C) are adapted from Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5).)*

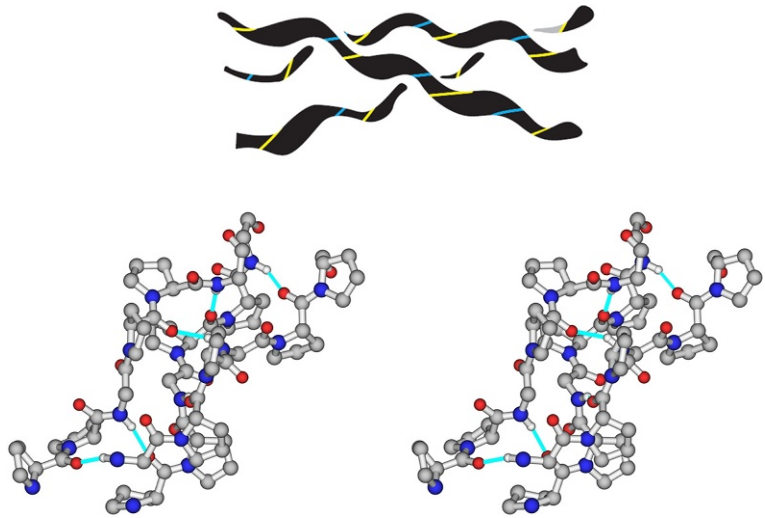


FIG. 7.9 *Top*: the overall view of poly(Pro)II, the right-handed superhelix composed of three left-handed helices. *Below*: the collagen triple superhelix (stereo drawing). Each of these three chains consists of Gly-Pro-Pro repeats. H-bonds connect NH-groups of Gly to C′O-groups of the first Pro in the Gly-Pro-Pro triplet: NH group of Gly of chain "1" is bonded to C′O group of Pro of chain "2," C′O group of Pro of chain "1" is bonded to NH group of Gly of chain "3."

TABLE 7.1 The Main Geometric Parameters of the Most Abundant Secondary Structures in Proteins

Structure	H-Bonding	Residues Per Turn	Shift Per Residue (Å)	φ (°)	ψ (°)
Helix α_R	CO_0–HN_{+4}	+3.6	1.5	−60	−45
Helix $(3_{10})_R$	CO_0–HN_{+3}	+3.0	2.0	−50	−25
Sheet β↑↓	Between chains[a]	−2.3	3.4	−135	+150
Sheet β↑↑	Between chains[a]	−2.3	3.2	−120	+135
Helix poly(Pro)II	No	−3.0	3.0	−80	+155

Data are from Schulz and Schirmer (1979, 2013) and Creighton (1993). All values are approximate. In the "Residues per turn" column "+" denotes the right-handed helix, and "−" the left-handed.
[a]*The distance between strands in the β-sheet is 4.8 Å.*

in antiparallel β-hairpins). The appearances of most typical β-bends and conformations of their constituent residues are presented in Fig. 7.10. Compare Fig. 7.10C with Fig. 7.4 and Table 7.1 and pay attention to the fact that conformations of turn I (and especially of turn III) are close to that of the turn of a 3_{10}-helix.

Usually, bends comprise about half of the residues uninvolved in the regular secondary structures of the protein.

Another kind of irregular secondary structure is the β-bulge (Richardson et al., 1978) (Fig. 7.11). It is formed by a residue (or sometimes by a few residues) "inserted" in the β-strand and having a non-β-structural conformation. The bulge is typical only for edge strands of a β-sheet (usually of an antiparallel β-sheet); it increases the twist of this sheet and bends it.

A chain without any distinct structure is usually called a "coil."

When applied to proteins, the term "coil" covers pieces of chains without α and β structures but with a great many various particular conformations. These irregular pieces include β-turns and short portions of polyPro helices (Adzhubei et al., 1987).

The same term is also applied to unfolded protein chains (ie, those in strong denaturants; Nozaki and Tanford, 1967; Tanford et al., 1967; Lapanje and Tanford, 1967). Here, it means regions without any regular secondary structure and long-range order in the chain. However, some weak short range order (within a few consecutive residues only) cannot be ruled out in this "coil" (often called a "random coil").

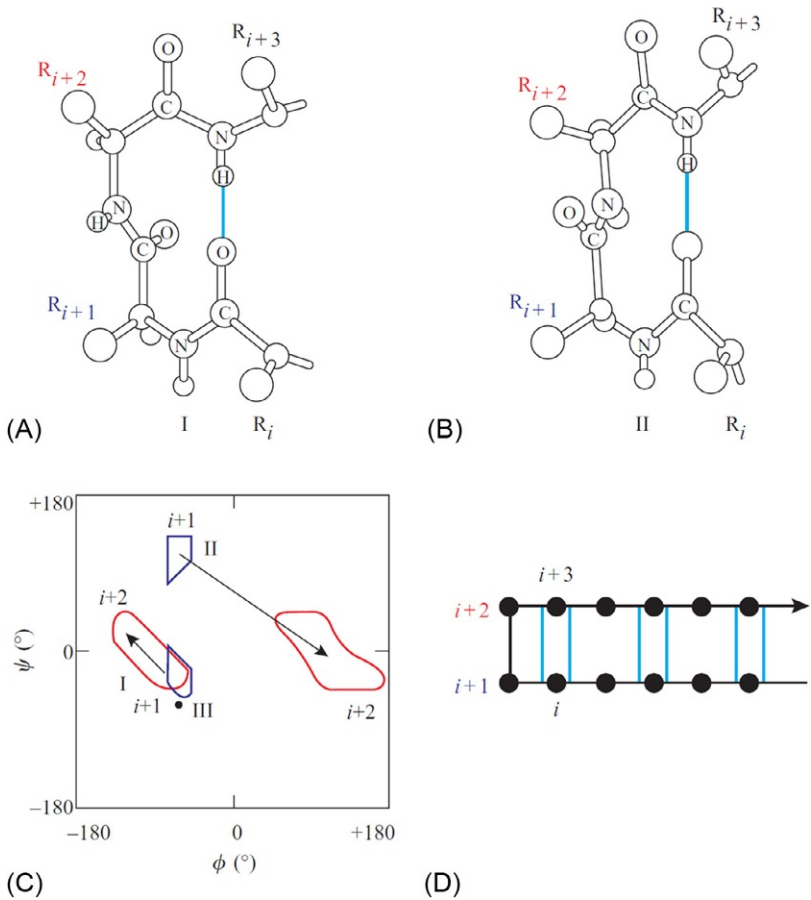


FIG. 7.10 β-turns. (A) The β-turn of type I (type III looks very similar and therefore is not given here). (B) The β-turn of type II. It differs from the β-turn I mainly by the inverted peptide group between the residues $i+1$ and $i+2$. (C) Conformations of the residues $i+1$ and $i+2$ are fixed by the H-bond closing the β-turns. In the β-turn III both these residues have the same conformation (denoted with a *bold dot*). The conformations of residues i and $i+3$ are not fixed in β-turns; they are fixed by the β-structure, when this structure extends the turn as shown in (D), which sketches a β-hairpin with the β-turn at its *top*. H-bonds are shown in *light-blue*. *(Parts (A), (B), and (C) are adapted from Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5).)*

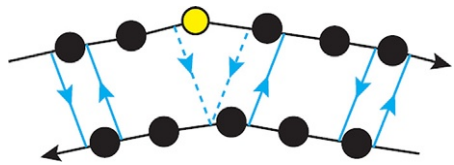


FIG. 7.11 The β-bulge in an antiparallel β-structure. Hydrogen bonds are shown as *blue arrows* directed from N–H to O═C groups. All residues (shown as *circles*) have β-structural conformations, except for that shown as a *yellow circle*, whose conformation is often nearly α-helical. One or even both H-bonds adjacent to this residue (shown by *broken arrows*) may be broken.

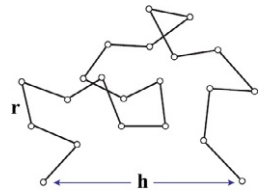


FIG. 7.12 The freely joint ("random-flight") chain: the simplest model of a random coil.

The most interesting features of the random coil (experimentally observed by using hydrodynamic techniques and light- and X-ray scattering (Tanford et al., 1967; Lapanje and Tanford, 1967) are its extremely low density and large volume, and a most peculiar dependence of its radius and volume on the chain length.

To shed light upon this peculiarity, let us consider the simplest model of a random coil, the so-called "loose joint chain" (Birshtein and Ptitsyn, 1966; Flory, 1969; Volkenstein, 1977), see Fig. 7.12. Its "links" are represented as sticks (each link can include a few chain monomers); the main distinctive feature of this model is that each stick can freely turn on the joint about the neighboring sticks. Let us assume that there are M sticks in the chain, and the length of each stick is r.

Such a chain can be described as a sequence of vectors $\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_M$ (may I remind you that in math a vector is a directed straight segment). These vectors have identical lengths r, and each of them is directed from the previous joint to the next one.

The sum of these vectors, $\mathbf{h} = \sum_{i=1}^{M} \mathbf{r}_i$, is just a vector running from the beginning of the chain to its end.

The average over all possible thermal fluctuations vector $\langle \mathbf{h} \rangle = 0$, because each vector $\mathbf{r}_i$ can have any direction.

Now let us find $\langle \mathbf{h}^2 \rangle$, the average value of $\mathbf{h}^2$; ie, let us have $\mathbf{h}^2$ averaged over all possible thermal fluctuations of the chain conformation (this averaging is denoted by $\langle\ \rangle$). The squared length of $\mathbf{h}$ is

$$\mathbf{h}^2 = \left(\sum_{i=1}^{M} \mathbf{r}_i \right)^2 = \sum_{i=1}^{M} \mathbf{r}_i^2 + \sum_{i=1}^{M} \sum_{j=1,\, j \neq i}^{M} \mathbf{r}_i \mathbf{r}_j \tag{7.1}$$

When averaging $\mathbf{h}^2$, we have to keep in mind that $\langle \mathbf{r}_i^2 \rangle$, t average squared vector $\mathbf{r}_i$, is just r^2, and the average scalar product of any vectors $\langle \mathbf{r}_i \mathbf{r}_j \rangle$ is *zero* when $i \neq j$, since free rotation of sticks provides equal probabilities of any direction of these vectors.

Hence, the average squared distance between the ends of a loose joint chain is:

$$\langle \mathbf{h}^2 \rangle = \left\langle \sum_{i=1}^{M} \mathbf{r}_i^2 \right\rangle + \left\langle \sum_{i=1}^{M} \sum_{j=1, j \neq 1}^{M} \mathbf{r}_i \mathbf{r}_j \right\rangle = \sum_{i=1}^{M} \langle \mathbf{r}_i^2 \rangle + \sum_{i=1}^{M} \sum_{j=1, j \neq 1}^{M} \langle \mathbf{r}_i \mathbf{r}_j \rangle = Mr^2 \tag{7.2}$$

ie, the linear dimensions (radius, etc.) of the coil increase with increasing number of chain links M as $M^{1/2}$. Consequently, the coil volume is proportional to $M^{3/2}$ although the volume of all "normal" (ie, fixed density) bodies increases only as the number of particles M, that is, much more slowly than $M^{3/2}$. This abnormally strong dependence of the coil volume on the chain length is the most prominent characteristic feature of the random coil. Specifically, this feature is responsible for the extremely low density of a coil formed by long chains, and consequently, for nearly zero contacts between distant links in the chain.

In addition, since the coil volume is proportional to $M^{3/2}$ and the probability that the chain ends meet is inversely proportional to the volume occupied by the coil, the probability of chain ends meeting is proportional to $M^{-3/2}$. This means that the free energy of loop closing in a coil-like chain increases with its length as $kT \times (3/2) \ln(M)$ (Flory, 1969), where $k \times (3/2) \ln(M)$ is the Flory's entropy of loop closure.

The free joint model presents a far too ideal coil in which any link may rotate by any angle, which is not actually true. However, Eq. (7.2) can be generalized as:

$$\left\langle \mathbf{h}^2 \right\rangle = Lr \tag{7.3}$$

where $L = Mr$ is the full ("contour") length of the chain proportional to the number of its constituent links, and r is the effective distance between the chain "free joints," ie, the characteristic length after which the chain "forgets" its direction (*note*: in a polypeptide chain, this characteristic length, also called the "length of the Kuhn segment," is 30–40 Å (when the polypeptide includes neither too flexible Gly nor too rigid Pro), ie, the Kuhn segment includes about 10 amino acid residues (Flory, 1969).

The equation for the coil size presented as Eq. (7.3) is general enough to be applied in the description of real polymers.

The more realistic "freely-rotating" model of the random coil (Fig. 7.13) gives the following estimate of the Kuhn segment length r (Birshtein and Ptitsyn, 1966; Flory, 1969):

$$r = l \frac{1 + \langle \cos\ \alpha \rangle}{1 - \langle \cos\ \alpha \rangle} \tag{7.4}$$

where l is the distance between the points where the chain turns (ie, the covalent bond length) and $\langle \cos \rangle$ is the average cosine of the turn angle (see Fig. 7.13).

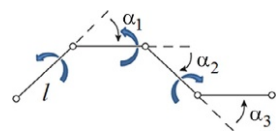


FIG. 7.13 The freely-rotating model of the random coil. Here, the "links" are again represented as sticks of equal lengths l; each stick can freely rotate about the axis formed by the preceding stick; α_1, α_2, … are angles between adjacent sticks.

In conclusion, a few words on how the secondary structure is determined by experiment. Of course, with X-ray (or accurate multidimensional NMR (nuclear magnetic resonance)) protein structures available (Fig. 7.14), the secondary structure can be derived from atomic coordinates. However, this is hardly possible for disordered proteins or unfolded protein chains. Nevertheless, by detecting closely positioned H-atom nuclei (with <4–5 Å between them), NMR reveals the secondary structure (mainly, α and β) even in these cases (Serdyuk et al., 2007).

NMR spectroscopy is based on applying radiowaves to excite the magnetic moments of nuclei aligned in a strong magnetic field. These nuclei must have an odd number of nucleons (protons and neutrons): then they have a magnetic moment, or spin. In proteins, these are natural "light" hydrogens (^{1}H), as well as introduced isotopes (^{13}C, ^{15}N, etc.). A key role in structure determination is played by NOESY (nuclear Overhauser effect spectroscopy). The magnetic resonance occurs at a radio frequency typical (in the given magnetic field) of the nucleus in question and slightly modified by its neighbors in chemical bonds and in space (which helps us to understand which atom of which residue is excited) (Serdyuk et al., 2007). The excitation can be propagated from the initial nucleus to a neighboring one (if it has a magnetic moment); the recipient will report on its excitation at its own frequency, thereby demonstrating the closeness of the two nuclei.

The characteristic feature of α-helices is the closeness of the H atom of the $C^{\alpha}H$-group to that of the NH-group of the fourth residue down the chain (towards the C-terminus), while the typical feature of the β-structure is closeness between H atoms of NH- and $C^{\alpha}H$-groups pertaining to immediate neighbors in the chain and to H-bonded residues in the β-sheet (Figs. 7.14B and 7.15).

However, the most important role in determining the secondary structure (mainly, α and β) is played by circular dichroism (CD) (Greenfield and Fasman, 1969; Creighton, 1993; Serdyuk et al., 2007).

For CD, a knowledge of the overall spatial structure of the protein is not required. On the contrary, structural studies of a protein are usually started with CD. This method is based on differing absorptions of clockwise and

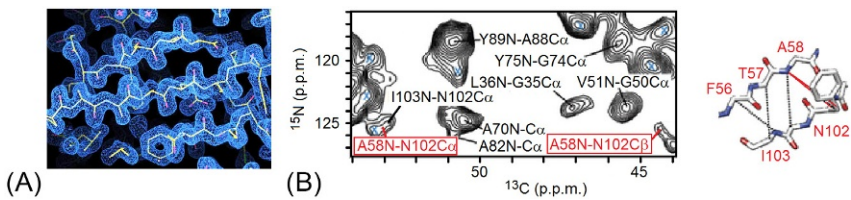


FIG. 7.14 (A) Region of electron density map with the β-structure. (B) A part of the 2D NOESY spectrum; two cross-peaks corresponding to a portion of the β-structure (shown on the *right*) are marked in *red*. *(The images are taken from http://learn.crystallography.org.uk/wp-content/uploads/2014/01/electronDensity.jpg and http://www.nature.com/nmeth/journal/v9/n12/images/nmeth.2248-F1.jpg, respectively.)*

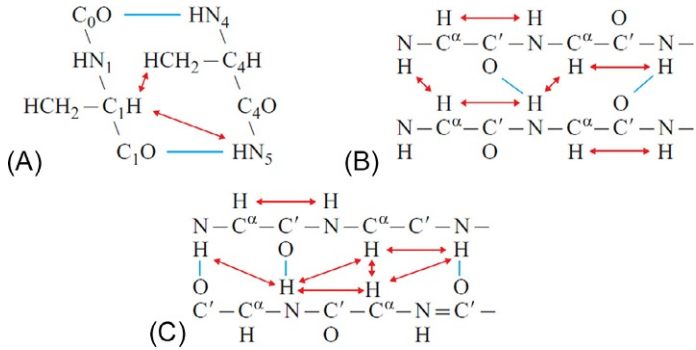


FIG. 7.15 Approach (↔) of the nuclei of H-atoms characteristic of the α-helix (A), and of the parallel (B) and antiparallel (C) β-structure. In (A), indices at atoms of the main chain indicate the relative location of residues in the chain.

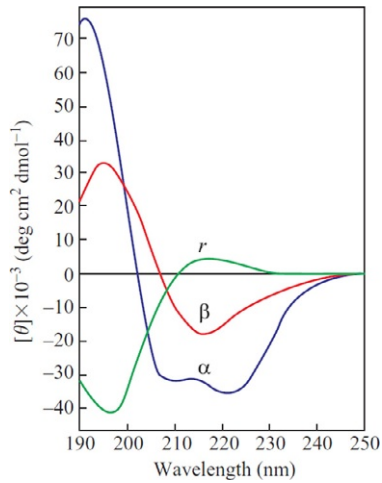


FIG. 7.16 Typical far UV CD spectra for polylysine as the α-helix (α), β-structure (β), and random coil (r). *(Adapted from Greenfield, N.J., Fasman, G.D., 1969. Computed circular dichroism spectra for the evaluation of protein conformation. Biochemistry 8, 4108–4116.)*

counterclockwise polarized light, which are caused specifically by helices of different handedness. Owing to this difference, plane-polarized light turns into elliptically polarized light.

The typical ellipticity spectra for the "far UV region" (190–240 nm) are given in Fig. 7.16. These spectra depend on the asymmetry of the peptide group environment, and therefore report on the secondary structure.

The optical excitation of peptide groups in the far UV region occurs at a wavelength of about 200 nm. This wavelength is approximately twice as large as that required for excitation of separate atoms. The possibility to excite a

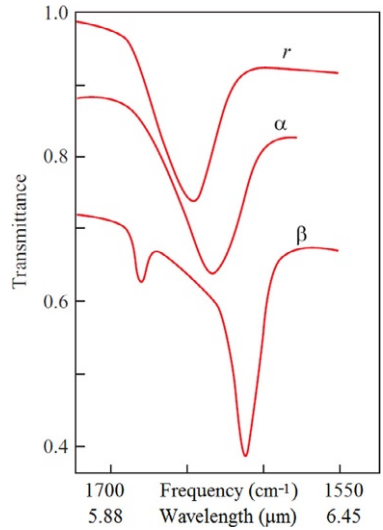


FIG. 7.17 Typical infrared (IR) transmittance spectra measured in heavy water (D_2O) for polylysine as the α-helix (α), β-structure (β), and random coil (*r*). The measurements were taken in the "amide I" region, reflecting vibrations of the C═O bond. *(Adapted from Susi, H., 1972. Infrared spectroscopy—conformation. Methods Enzymol. 26, 455–472.)*

peptide group with light of lower frequency is explained by delocalization of electrons over the group, which has been discussed previously.

In aromatic side groups, delocalization of electrons is still greater: here, they are "spread over" six atoms (while in peptide groups over three atoms). The CD spectra for aromatic groups fall within the wavelengths of about 250–280 nm (although their "tails" can reach even 220 nm). In this range of 250–280 nm (in the "near-UV region") the asymmetry of aromatic side-chain environments is studied, ie, the effects characteristic of not the secondary but the tertiary structure of the protein (Creighton, 1993; Serdyuk et al., 2007).

In passing, I could add that when the electron is delocalized still more (ie, in larger molecules with double partial bonds) its excitation moves from UV to visible (400–600 nm) light, and such molecules become dyes.

Apart from UV spectra, IR spectra can be exploited to reveal the secondary structure of polypeptides and proteins. They reflect the difference in vibration of peptide groups involved and uninvolved in various secondary structures (Susi, 1972; Creighton, 1993; Serdyuk et al., 2007) (Fig. 7.17). These measurements are more complicated than study of UV spectra, since the absorption range of ordinary water (H_2O) is nearly the same; therefore, heavy water (D_2O) is used. Also, these measurements, compared with UV ones, require more protein to be used and a higher protein concentration in solution.

REFERENCES

Adzhubei, A.A., Eisenmenger, F., Tumanyan, V.G., Zinke, M., Brodzinski, S., Esipova, N.G., 1987. Third type of secondary structure: noncooperative mobile conformation. Protein Data Bank analysis. Biochem. Biophys. Res. Commun. 146, 934–938.

Birshtein, T.M., Ptitsyn, O.B., 1966. Conformations of Macromolecules. Interscience Publishers, New York (Chapters 4, 9).

Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second ed. Garland, New York (Chapter 2).

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman, New York (Part 1, Chapters 2, 5).

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

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman, New York (Chapter 5).

Flory, P.J., 1969. Statistical Mechanics of Chain Molecules. Interscience Publishers, New York (Chapters 1–3, 7–8; Appendix G).

Greenfield, N.J., Fasman, G.D., 1969. Computed circular dichroism spectra for the evaluation of protein conformation. Biochemistry 8, 4108–4116.

Lapanje, S., Tanford, C., 1967. Proteins as random coils. IV. Osmotic pressures, second virial coefficients, and unperturbed dimensions in 6 M guanidine hydrochloride. J. Am. Chem. Soc. 89, 5030–5033.

Nozaki, Y., Tanford, C., 1967. Proteins as random coils. II. Hydrogen ion titration curve of ribonuclease in 6 M guanidine hydrochloride. J. Am. Chem. Soc. 89, 742–749.

Pauling, L., Corey, R.B., 1951a. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. Proc. Natl. Acad. Sci. U. S. A. 37, 235–240.

Pauling, L., Corey, R.B., 1951b. The pleated sheet, a new layer configuration of polypeptide chains. Proc. Natl. Acad. Sci. U. S. A. 37, 251–256.

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

Ptitsyn, O.B., Finkel'shtein, A.V., 1970. Predicting the spiral portions of globular proteins from their primary structure. Dokl. Akad. Nauk SSSR (in Russian) 195, 221–224.

Ramachandran, G.N., Sasisekharan, V., 1968. Conformation of polypeptides and proteins. Adv. Protein Chem. 23, 283–438.

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

Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5).

Serdyuk, I.N., Zaccai, N.R., Zaccai, J., 2007. Methods in Molecular Biophysics: Structure, Dynamics, Function. Cambridge University Press, Cambridge (Chapters H1–H3, J1–J2).

Susi, H., 1972. Infrared spectroscopy—conformation. Methods Enzymol. 26, 455–472.

Tanford, C., Kawahara, K., Lapanje, S., Hooker Jr., T.M., Zarlengo, M.H., Salahuddin, A., Aune, K.C., Takagi, T., 1967. Proteins as random coils. III. Optical rotatory dispersion in 6 M guanidine hydrochloride. J. Am. Chem. Soc. 89, 5023–5029.

Traub, W., Piez, K.A., 1971. The chemistry and structure of collagen. Adv. Protein Chem. 25, 243–352.

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

Volkenstein, M.V., 1977. Molecular Biophysics. Academic Press, New York (Chapters 4, 6).

This page intentionally left blank

Lecture 8

At this point, we could pass on to the formation and decay of the secondary structure. However, prior to that I would like to talk about the fundamentals of statistical physics, thermodynamics, and kinetics "in general," inasmuch as otherwise it is difficult to discuss the stability of the secondary structure, the stability of proteins, cooperative transitions in polypeptides and proteins and the kinetics of these transitions.

Thermodynamics provides an idea of the types of possible cooperative transitions in systems incorporating a great many particles. Statistical physics advises on when and what transitions may occur in the system of particles in question and gives details of these transitions based on the properties of the studied particles and their interactions.

First of all, we will consider the main concepts of statistical physics and thermodynamics, namely, entropy, temperature, free energy and partition function. In doing this, I will follow the generally recognized books *The Feynman Lectures on Physics* (Feynman et al., 1963) and *Statistical Physics* by Landau and Lifshitz (1980).

Systems with numerous degrees of freedom (ie, comprising a lot of molecules or even just one large and flexible molecule) are described by means of statistical physics. It is "statistical" because such a large system has zillions of configurations. Here is an illustrative example. If each of N links of a chain may have only two configurations (eg, "helical" and "extended" ones), then the whole N-link chain has 2^N possible configurations. In other words, a "normal" 100-link protein chain may have at least 2^{100} configurations, that is, about 1 000 000 000 000 000 000 000 000 000 000 of them! This is an enormous number. Consideration of all of them at a rate of 1 nanosecond (10^{-9} s) per configuration would take 3×10^{13} years, that is more than 1000 lifetimes *of the Universe*… And in the experimental tube, there are billions of such chains, not to mention the solvent. If we had been going to consider all their configurations, we would have been lost forever. Certainly, we are interested in much more simple and reasonable things, such as the average (ie, *statistically averaged*) helicity of chains and its change upon heating. A peculiar feature of statistical averaging (ie, neglecting all minor details) is a crucial simplification of events.

In statistical averaging, the major role is played by *entropy*. It shows how many configurations (in other words, *microstates*) of the system correspond to its observed *macrostate* (averaged by the observation time and the number of molecules studied). We considered a similar example earlier, though it was only a special case with the number of *microstates* in space for a molecule limited by a volume, and the entropy of the molecule proportional to the

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00008-9

logarithm of this volume (and the molecule's being in the given volume, eg, "in this room," was the "macrostate" of the molecule).

Here, it would be timely to answer the question that is only natural: Why do physicists prefer to consider the *logarithm* of microstate number but not the number itself? The answer is that in considering many separate systems (eg, separate molecules) we have to sum up their energies and degrees of freedom, while the numbers of accessible states are to be multiplied (if one molecule has ten microstates and another as many, these two molecules have 100 different combinations of microstates). This makes the calculations inconvenient and too bulky. But *logarithms* of the numbers of accessible states are to be summed up (you remember that $\ln(AB) = \ln(A) + \ln(B)$), like energies or volumes. Therefore, the logarithms are convenient for calculations. Moreover, the additivity of logarithms allows us to use the potent differential calculus.

Now let us talk about *temperature*, or rather, *absolute temperature*.

First: Why, from the times of Clapeyron and Kelvin, is it counted off $-273\ ^{\circ}C$?

Because (unlike $0\ ^{\circ}C$, the melting point of one of a great many crystals) this point is *universal*: it is obtained by extrapolation to zero of $PV =$ pressure $\times$ volume of *all* quantities of *all* gases (Fig. 8.1).

Inner voice: But these extrapolations ignore all visible deviations of experimental points from the straight lines that extrapolate only the linear middles of the *PV*-on-temperature dependencies!

Lecturer: Exactly! Such deviations are connected with either the gas to liquid transition (at low temperatures) or with dissociation of gas molecules (at too high temperatures). And without *ignoring* these "side-effects," a universal law would be never derived! Note that blind adherence to experiment can prevent from discovering a universal gas law, which can be screened by various "side-effects" such as gas condensation, etc., of various substances…

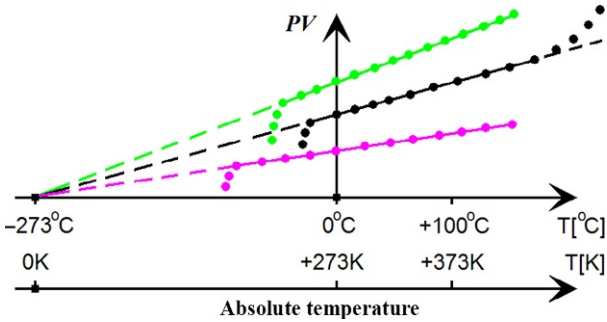
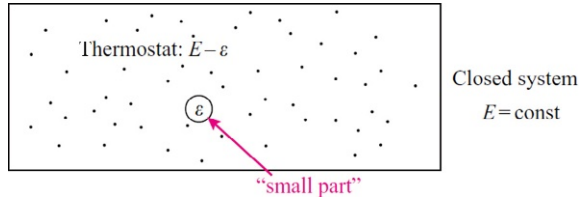


FIG. 8.1 A scheme of determination of the zero point of absolute temperature from extrapolation to zero of $PV =$ pressure $\times$ volume of various quantities of various gases at various (but not too high) pressures. Dots represent "experimental" points, dashed lines show extrapolations of linear parts of the *PV*-on-temperature dependencies.

Second: It is to be shown that *temperature* (which is now only a phenomenological concept) is closely connected with entropy: *where there's no numerous states (no entropy), there's no temperature either*.

To clarify this connection, let us follow J.W. Gibbs and consider a closed system (with no energy exchange with the environment). Let its total energy be E, and let this system be in equilibrium, that is, all its microstates with the energy E are equally probable (and those of different energy $E' \neq E$ have zero probability).

Let us pick out an "observed small part" of the system (eg, a molecule in gas or a macromolecule with its liquid environment). Then the rest of the system may be regarded as a thermostat in which the "small part" is immersed.



Let us divide all the system's microstates having the total energy E into such classes that each of them corresponds to one microstate of the chosen "small part." The more microstates incorporated in a given class, the higher probability of observation of this class, that is, the higher the probability of observation of a certain state of our "small part."

Let a microstate of our "small part" be given (eg, let a molecule in gas have a certain position in space and a certain speed). Let its energy be ε. Since the system ("small part"+"thermostat") is closed, *its total energy E is conserved* (as stated by the energy conservation law), and thus the thermostat energy is $E-\varepsilon$. Let as many as $M_{\text{therm}}(E-\varepsilon)$ of thermostat microstates correspond to this energy. Then the probability of observation of this state of our "small part" is simply proportional to $M_{\text{therm}}(E-\varepsilon)$.

Note: Here, an implicit assumption has been made that a certain microstate of the system produces no effect on thermostat microstates. Strictly speaking, this is not quite true (or rather, this is true only for an ideal gas as the thermostat), but taking into consideration all events at the border of "our system" and the thermostat would obscure the entire narration. So, to combine strictness and clearness, let us assume for the time being that our "observed part" is encapsulated and thereby separated from thermostat molecules; if required (not in these lectures), the interaction between the "observed part" and the thermostat can be considered separately.

With the number of the thermostat states equal to $M_{\text{therm}}(E-\varepsilon)$, its logarithm *by definition* is proportional to the thermostat entropy:

$$S_{\text{therm}}(E-\varepsilon) = \kappa \ln \left[M_{\text{therm}}(E-\varepsilon)\right] \tag{8.1}$$

The coefficient κ is used here only to have the entropy measured in cal K^{-1}, as usual; as you will see later, it appears to be simply the Boltzmann constant.

The energy of the "small part" ε must be relatively small as well. Therefore, we can use an ordinary differential expansion of $S_{\text{therm}}(E-\varepsilon)$ over the small parameter ε (you may remember that $f(x_0+\mathrm{d}x)=f(x_0)+\mathrm{d}x\ (\mathrm{d}f/\mathrm{d}x)|_0+\frac{1}{2}\ (\mathrm{d}x)^2\ (\mathrm{d}^2f/\mathrm{d}x^2)|_0+\ldots=f(x_0)+\mathrm{d}x\ (\mathrm{d}f/\mathrm{d}x)|_0$ at a small $\mathrm{d}x$ with $(\mathrm{d}f/\mathrm{d}x)|_0$ meaning that the derivative $\mathrm{d}f/\mathrm{d}x$ is taken at the point x_0). So,

$$S_{\text{therm}}(E-\varepsilon)=S_{\text{therm}}(E)-\varepsilon\times(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E \tag{8.2}$$

Note: Since both S and E are proportional to the number of particles, $\mathrm{d}S_{\text{therm}}/\mathrm{d}E$ is independent of the number of particles in the thermostat, while $\mathrm{d}^2S_{\text{therm}}/\mathrm{d}E^2$ is *inversely* proportional to this number, that is, $\mathrm{d}^2S_{\text{therm}}/\mathrm{d}E^2\to 0$ in a very large thermostat; this allows us to neglect members of the order of ε^2 (as well as ε^3, etc.) in Eq. (8.2).

Thus, the number of accessible thermostat microstates depends on the energy ε of our "small part" as

$$\begin{aligned} M(E-\varepsilon)&=\exp\left[\frac{S_{\text{therm}}(E-\varepsilon)}{\kappa}\right]\\ &=\exp\left[\frac{S_{\text{therm}}(E)}{\kappa}\right]\times\exp\left\{-\varepsilon\left[\frac{(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E}{\kappa}\right]\right\} \end{aligned} \tag{8.3}$$

Here, neither the common multiplier $\exp[S_{\text{therm}}(E)/\kappa]=M(E)$ nor the number $(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E$ depends on ε or on a concrete microstate of our "small part" in general.

Since the number of microstates must increase with increasing energy (the higher the energy, the greater the number of ways it can be divided), Eq. (8.3) expresses the following simple idea: the more energy taken from the thermostat by our "small part," the less energy kept by the thermostat, and hence, the smaller the number of ways it can be divided. Moreover, this equation shows that the decrease of the number of accessible thermostat microstates (the number of ways to divide its energy) depends *exponentially* on the energy of our "small part."

Conclusion: The probability of observation of a certain microstate of our "small part" (molecule, etc.) is proportional to $\exp[-\varepsilon\{(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E/\kappa\}]$, where ε is the energy of this "small part," and the magnitude $\{(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E/\kappa\}$ depends *not* on the "small part" but only on averaged features of its environment.

But according to Boltzmann (=Boltzmann–Gibbs) distribution, the probability of a molecule keeping a certain state with energy ε is proportional to $\exp(-\varepsilon/k_{\text{B}}T)$ (where T is temperature, and k_{B} is the Boltzmann constant). A comparison of the identical expressions $\exp[-\varepsilon\{(\mathrm{d}S_{\text{therm}}/\mathrm{d}E)|_E/\kappa\}]$ and $\exp(-\varepsilon/k_{\text{B}}T)$ yields

$$(dS_{\text{therm}}/dE)|_E = \frac{1}{T} \tag{8.4}$$

and κ in Eqs. (8.1) and (8.3) turns out to be the Boltzmann constant (k_B), provided the energy is measured (as usual) in Joules (J) (or in calories, cal), temperature in K, and entropy in J K^{-1} (or in cal K^{-1}).

Eqs. (8.1) and (8.4) are the main equations of statistical physics and thermodynamics: they define the temperature as the reciprocal of the rate of entropy (or of the logarithm of the number of microstates) change with the system energy E.

In particular, they show that $\ln[M(E+k_BT)] = S(E+k_BT)/k_B = [S(E)+(k_BT)(1/T)]/k_B = \ln[M(E)]+1$, that is, an energy increase by k_BT results in an "e" (=2.72)-fold (approximately three-fold) increase of the number of microstates, *independently* of the system size, its inner forces, etc.

They also allow us to find the corresponding temperature value for each energy of any large system (thermostat), provided we know the number of its microstates at different energies (also called the "energy spectrum density"), or rather, the dependence of the logarithm of the energy spectrum density on this energy. The schematic diagram is given in Fig. 8.2.

It is essential that the microstates are highly numerous because the derivative dS/dE can be taken only when a small (as compared with kBT) energy interval houses many microstates. That is why temperature appears only in sciences dealing with enormous numbers of accessible states.

Let us continue considering a small system in the thermostat whose temperature is equal to T. Eqs. (8.1) and (8.3) show that for our system the probability of being in a given state i with the energy εi at temperature T is

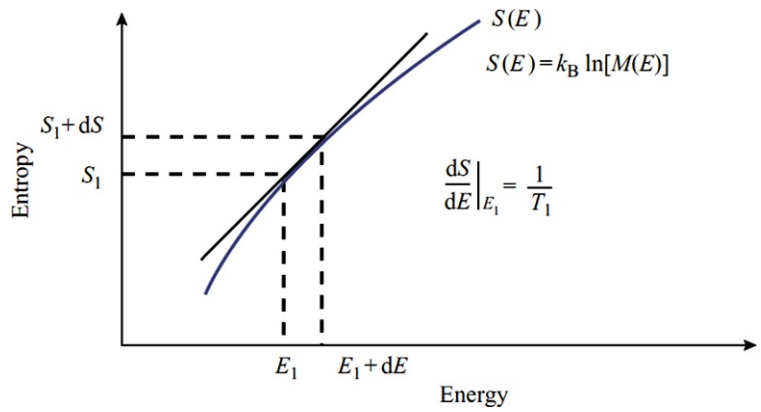


FIG. 8.2 Determination of the temperature of a large system. The bold curve shows the dependence of entropy S on the system energy E. The curve slope, dS/dE, determines the temperature T corresponding to this energy E. $M(E)$ is the number of microstates with the energy E, that is, the density of the energy spectrum of the system.

$$w_i(T) = \frac{\exp(-\varepsilon_i/k_B T)}{Z(T)} \tag{8.5}$$

where

$$Z(T) = \sum\nolimits_j \exp\left(-\varepsilon_j/k_B T\right) \tag{8.6}$$

is the normalization factor which takes into account that the sum of probabilities of all states, $\sum_j w_j$ is necessarily equal to 1 (here and above the sum $\sum_j$ is taken over all microstates j of the studied "small system").

The value Z is called *partition function* for the studied system. Provided Z is known, Eq. (8.5) allows us to calculate the probability w of each microstate of this system at a given temperature. Then the average energy of the system at this temperature

$$E(T) = \sum\nolimits_j w_j \varepsilon_j \tag{8.7}$$

and its average entropy

$$S(T) = k_B \sum\nolimits_j w_j \ln\left(1/w_j\right) \tag{8.8}$$

Note that Eq. (8.8) averages $\ln(1/w_j)$ over all microstates j of the system, allowing for their probabilities w_j.This equation provides a direct generalization of the determination of entropy $S = k_B \ln[M(E)]$ already familiar to us (see Eq. 8.1). This is a generalization of averaging for the case when w_j have more than only two values that follow from the energy conservation law: $w_j = 1/M(E)$ for all $M(E)$ states where $E_j = E$ and $w_j = 0$ when $Ej \neq E$.

Inner voice: I feel that the meaning of Eq. (8.8) must be explained better, and that the term $S(T)/k_B$ should be proved to be the logarithm of the average number of system states.

Lecturer: Then, please be tolerant and we will go into some maths…

Following J.W. Gibbs, let us consider a large number N of equal systems each of which may be in the state "1" with the probability w_1, in the state "2" with the probability w_2, …, in the state "J" with the probability w_J. Then on average, among N systems considered, those in the state "1" amount to $n_1 = w_1 N$, those in the state "2" amount to $n_2 = w_2 N$, and so on (while $\sum_j w_j N \equiv \sum_j n_j = N$).

$\frac{N!}{n_1! \times n_2! \times \ldots \times n_J!}$ is the well-known number of different ways to divide N systems yielding n_1 of them having state "1", n_2 having state "2",…, n_J having state "J" (here, $N! \equiv N \times (N-1) \times \ldots \times 2 \times 1$ is the number of different enumerations of the N systems, $n_1!$ is the number of enumerations of n_1 systems having the identical state "1", etc.). Now we use Stirling's approximation $n! \approx (n/e)^n$ and obtain $\frac{N!}{n_1! \times n_2! \times \ldots \times n_J!} \approx \prod_j \left(\frac{N}{n_j}\right)^{n_j} = \left[\prod_j \left(\frac{1}{w_j}\right)^{w_j}\right]^N$.

And since this number (the total number of states for N independent

systems) is simply the number of states of one system to the Nth power, the average number of states of *one* system is equal to $\prod_j \left(\frac{1}{w_j}\right)^{w_j}$, and its logarithm is (as it has been expected to be) the term $S(T)/k_B$ of Eq. (8.8). That's all.

The partition function (which may seem to be simply the normalization coefficient in Eq. (5)) plays a most important role in statistical physics because the quantity $Z(T)$ allows direct calculation of the free energy of a system enclosed in a fixed volume,

$$F(T) = E(T) - TS(T) = \sum\nolimits_j w_j \left\{ \varepsilon_j - T\left[-k_B \ln\left(w_j\right) \right] \right\} = -k_B T \ln\left[Z(T)\right] \quad (8.9)$$

(here we used Eqs. (8.7) and (8.8) and then Eq. (8.5)).

Derivatives of F determine all other thermodynamic functions:

$$S(T) = -\mathrm{d}F/\mathrm{d}T$$

We have already seen this equation; please check it up by yourselves using Eqs. (8.9), (8.6), (8.5), (8.8); and

$$E(T) \equiv F(T) + TS(T) = F(T) - T(\mathrm{d}F/\mathrm{d}T) \equiv \mathrm{d}(F/T)/\mathrm{d}(1/T).$$

Important secondary notes:

1. If the "small system" has many degrees of freedom, it has its own entropy and therefore its own (internal) temperature. The internal temperature of the "small system," T_{in}, is equal to the thermostat temperature T because, as follows from the above definitions and equations, $T_{in} \equiv \mathrm{d}E_{in}(T)/\mathrm{d}S_{in}(T) = \mathrm{d}[F_{in}(T) - T(\mathrm{d}F_{in}/\mathrm{d}T)]/\mathrm{d}(-\mathrm{d}F_{in}/\mathrm{d}T) = [\mathrm{d}F_{in}/\mathrm{d}T - \mathrm{T}(\mathrm{d}^2F_{in}/\mathrm{d}T^2) - \mathrm{d}F_{in}/\mathrm{d}T]\ \mathrm{d}T/(-\mathrm{d}^2F_{in}/\mathrm{d}T^2)\ \mathrm{d}T = T$.
2. The total energy incorporates kinetic and potential energies. The former depends on particle speeds only, while the latter on their positions in space, and not on the speeds. The "microstate" of each particle is determined by its coordinate in space and by its speed. In classical (not quantum) mechanics, any combinations of speeds and coordinates are allowed (as we know, Heisenberg's Quantum Uncertainty Principle, $\Delta v \Delta x \approx \hbar/m$, imposes restrictions on the speed–coordinate combinations, but at room temperature this is important for very light particles, ie, virtually electrons only). This means that probabilities for coordinates and speeds can be "uncoupled," that is, $w(\varepsilon_{kinet} + \varepsilon_{coord}) \sim \exp(-\varepsilon_{kinet}/k\mathrm{B}T) \times \exp(-\varepsilon_{coord}/k\mathrm{B}T)$. Further simple calculations (you can make them yourselves) will show that free energies, energies and entropies can also fall into kinetic and coordinate parts, that is, $F = F^{kinet} + F^{coord}$, etc. It is important that kinetic parts are *independent* of the system configuration and can be neglected when considering

conformational changes. Therefore, further on we will discuss *only* configurational (or "conformational") energy spectra, energies, entropies, etc.

3. Above, we summed over microstates, whereas in the frame of classical mechanics we can equally well integrate over coordinates and speeds that determine the microstate of each particle.
4. Equilibrium temperature must be *positive*. Otherwise, probability integration over speeds, that is, $\int \exp\left(-mv^2/2k_BT\right)dv$, turns into infinity at great speeds, and the system "explodes." Therefore, the stable state cannot be observed in those conformational energy spectrum regions where the spectrum density (and, hence, the entropy of the system) decreases with increasing energy: for these regions, $T<0$ (see Eqs. (8.4) and (8.1), and Fig. 8.3A).
5. The quantity k_BT is measured in units such as "energy per particle" or "energy per mole ($=6.02\times10^{23}$) of particles." Had temperature been expressed in energy units from the very outset, Boltzmann's constant k_B would never have been used at all. However, historically it happened that the "degree" was first introduced as a temperature unit, and then it became evident that it could be easily converted into something like "energy per particle" through multiplying by a certain (Boltzmann) constant. Accordingly, k_B is measured in the units "energy per particle per degree." Its numerical value depends on the energy unit used: "joule per particle," "calorie per mole of particles," etc. In accordance with the measured in joules energy cost of a "degree" (K), $k_B=1.38\times10^{-23}$ joule particle^{-1} K^{-1}. However, apart from "per particle," k_B can be calculated per mole ($=6.02\times10^{23}$) of particles. To do this we multiply and divide k_B by 6.02×10^{23}, and have: $k_B=1.38\times10^{-23}$ (joule/particle/degree) $=1.38\times10^{-23}$ (joule $\times$ [6.02×10^{23}]/[6.02×10^{23} particles]) degree^{-1} $=[1.38\times10^{-23}\times6.02\times10^{23}]$ (joule/mol of particles) degree^{-1} $=$ 8.31 (J/mol) degree^{-1} $=1.99$ (cal/mol) degree^{-1} (since 1 cal $=4.18$ J).

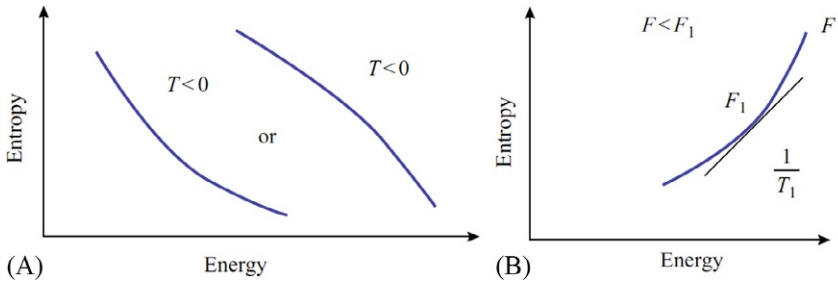


FIG. 8.3 Regions of the $S(E)$ plot that do not correspond with any stable state of the system. (A) Regions where entropy $S(E)$ decreases with increasing energy (E): here $T=1/(dS/dE)<0$. (B) A "concave" region of the $S(E)$ plot: here at the point of contact (where $dS/dE=1/T_1$) the free energy $F=E-TS$ is not lower but higher than that of neighboring sections of the plot $S(E)$; cf. Fig. 8.4.

The last value, 1.99 cal mol^{-1} K^{-1}, is traditionally referred to as the "gas constant" R.

6. Specific entropy is often measured in entropy units, "eu": eu = cal mol^{-1} K^{-1} = k_B/1.99 ≈ k_B/2. Hence, 1 eu corresponds to ≈ $e^{1/2}$ ≈ 1.65 states per particle of the system.

Now at last we can start considering *conformational changes*. It was in order to give them a competent consideration that we reviewed the basics of statistical physics and thermodynamics.

Conformational changes can be gradual or sharp (the latter are called "phase transitions"). Let us try and distinguish one kind from the other using energy spectrum density plots such as those presented in Fig. 8.2.

First of all, we have to consider *thermodynamics* and learn how to locate the stable state(s) of our system at a given temperature of the medium.

Let the number of states of our system (macromolecule) be $M(E)$ when its energy is E, and let the medium (thermostat) temperature be T_1. Given that we have the plot of entropy $S(E) = k_B \ln[M(E)]$ and know T_1, how can we find the plot's point corresponding to the stable state of our system? The temperature T_1 determines the plot's slope dS/dE at the sought point: here $dS/dE = 1/T_1$ since the temperature of our macromolecule is equal to the medium's temperature. Thus, the corresponding tangent to the curve $S(E)$ gives us the possibility of finding the point we are looking for.

However, there may be several such points with the given plot slope of $1/T_1$ (see Fig. 8.6). Which of them corresponds to the stable state? Let us consider the tangent at the point E_1 where $dS/dE|_{E1} = 1/T_1$. The equation describing such tangent is $S - S(E_1) = (E - E_1)/T_1$, or, what is the same, $E - T_1S = E_1 - T_1S(E_1)$. The value $F_1 = E_1 - T_1S(E_1)$ is simply the system's free energy at temperature T_1. As seen, along the tangent the value $E - T_1S$ is constant. Everywhere to the left of the tangent the value of $E - T_1S$ is lower and everywhere to its right higher than on the tangent itself (Fig. 8.4).

Specifically, the latter means that concave regions of $S(E)$ (Fig. 8.3B) cannot correspond to a stable state, since at the contact point, F is not lower but higher than that in the neighboring sections of the curve $S(E)$. The latter means that the system can decrease its free energy (ie, shift towards a more stable state) by moving from the contact point along the concave curve.

If the curve $S(E)$ is convex along its entire length, then its slope decreases with increasing energy E. Hence, each value of slope $1/T$ corresponds to only one point of the curve (Figs. 8.2 and 8.5), that is, this point corresponds to the sought point for the stable state at the given temperature T. As the temperature changes, this point gradually moves and the system gradually changes its entropy and energy (Fig. 8.5A and B) and its thermodynamic state (Fig. 8.5C).

If the $S(E)$ slope alternatively decreases and increases with increasing energy E (Fig. 8.6), then there may be several tangents with the same slope, and the contact point of the extreme left of these tangents (with the lowest F) reflects

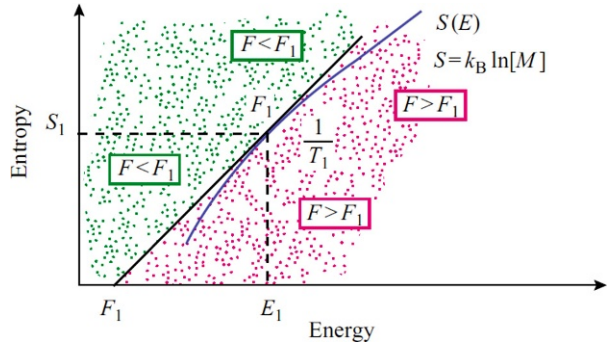


FIG. 8.4 Graphical definition of temperature and free energy. The bold curve $S(E)$ shows the dependence of entropy $S=k_B \ln(M)$ on the energy E. $M(E)$ is the energy spectrum density of the system's states. The slope of $S(E)$ determines the temperature T: $dS/dE=1/T$. Physically possible temperatures $T>0$ correspond to a rise in $S(E)$. F_1 is the free energy (corresponding to the given energy spectrum) at temperature T_1 (that determines the tangent to the $S(E)$ curve). The magnitude $F=E-T_1S$ is constant along the tangent and equal to $F_1=E_1-T_1S_1$. At the left and above the tangent, $F=E-T_1S<F_1$, while at its right and below, $F=E-T_1S>F_1$. Since here the curve $S(E)$ is convex, that is, it is below the tangent, the contact point corresponds to the minimum free energy at temperature T_1. For other explanations, see text.

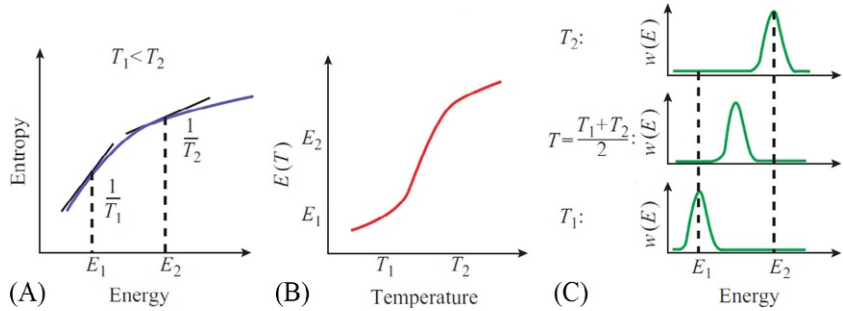


FIG. 8.5 A gradual change of the system state with changing temperature (from T_1 to T_2). The requirement is convexity of the curve $S(E)$. With $S(E)$ known (A), $T(E)$ and then $E(T)$ can be found (B). $w(E)$ (C) is the probability of having energy E at the given temperature T; $w(E)$ is proportional to $\exp[-(E-TS(E))/k_BT]$.

the most stable state. Up to a certain temperature T^*, the "best" tangent (with the lowest F) will be one in the region of small energies, whereas beyond T^* the "best" tangent will be found in the region of greater energies (and greater entropies).

At temperature T^*, structures with low and high energies will have equal free energies and equal probabilities of existence. This means, that among many identical systems at transition temperature T, half of them are in the low-energy state, while the other half is in the high-energy state (see Fig. 8.6C). In other

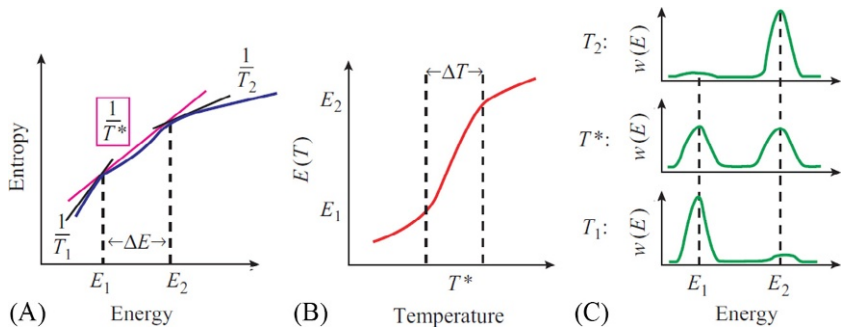


FIG. 8.6 Phase transition of the "all-or-none" type (in macroscopic bodies, it is called a "first-order phase transition") is characterized by a sharp change of the system's state with changing temperature. The requirement is a concave region on the curve $S(E)$ in (A): at the transition temperature T^*, the free energy at the center of this region is higher than at its flanks (this region corresponds to unstable states of the system, see Fig. 8.3b).The transition occurs within a narrow temperature range ($\Delta T = T_2 - T_1$) corresponding to the co-existence (ie, approximately equal probabilities) of low- and high-energy states. The tangents correspond to the mid-transition temperature T^*, as well as to $T_1 < T^*$ and $T_2 > T^*$. Note that temperature T^* of the "all-or-none" transition (ie, of the first-order phase transition) exactly coincides with the middle of the sharp energy change (B) and with maximum splitting of the distribution of probability $w(E)$ over the states (C).

words, half of the time each system is in the high-energy state, and for the other half it is in the low-energy state. The "co-existence" of two states of a system, equally probable but utterly different in energy, will occur within a certain very narrow (specifically, for large "macroscopic" systems) temperature range around T^*; we will estimate it soon.

It is of major importance, that the states with "medium" energies will not be displayed as probable states of the system, since, owing to the $S(E)$ plot concavity, the points for "intermediate" states lie on the right of the tangent corresponding to the transition temperature T^*. In other words, at temperature T^*, the free energy of these "intermediate" states is higher than that of structures corresponding to both contact points, and the probability of manifestation of the intermediates is nearly zero. Then the two stable states are said to be separated by a "free energy barrier."

These are conditions for the "all-or-none" transition.

In microscopic systems (in proteins, in particular), this transition can occur as a "jump" over the free energy barrier at the transition temperature T^*. We will discuss this later on.

In macroscopic systems (eg, in a tube with freezing water) the barrier (at temperature T^*) is so high that it would take almost infinity to overcome it. Therefore, macroscopic systems possess *hysteresis*, that is, they preserve their state up to slight overcooling (when freezing) or overheating (when melting) as compared with temperature T^*; after that the transition runs through a temporary (ie, unstable) phase separation in the system (eg, into liquid and

solid). Such a transition in macroscopic systems is called a *first-order phase transition*.

Now I would like to estimate the temperature interval corresponding to coexistence of low- and high-energy states of the system. In the middle of the transition, at temperature T^*, the free energy of the low-energy phase, $F_1(T^*) = E_1 - T^*S_1$, is equal to the free energy of the high-energy phase, $F_2(T^*) = E_2 - T^*S_2$. That is, in the middle of transition

$$E_2 - E_1 = T^*(S_2 - S_1) \tag{8.10}$$

At a small (δT) temperature deviation from T^*, the free energies of the phases change slightly. The difference between them amounts to:

$$\begin{aligned} \delta F &= F_1(T^* + \delta T) - F_2(T^* + \delta T) = [F_1(T^*) + (\mathrm{d}F_1/\mathrm{d}T)\delta T] \\ &\quad -[F_2(T^*) + (\mathrm{d}F_2/\mathrm{d}T)\delta T] = -S_1\delta T + S_2\delta T \\ &= \delta T(S_2 - S_1) \end{aligned} \tag{8.11}$$

The phases co-exist (ie, their probabilities are nearly equal: say, the probability ratio varies from 10:1 to 1:10) as long as $\exp(-\delta F/k\mathrm{B}\ T^*)$ is between about 10 and 1/10, that is, as long as $\delta F/k\mathrm{B}T^*$ is somewhere between $\ln(10) \approx +2$ and $\ln(1/10) \approx -2$. In this region δT is somewhere between $+2k\mathrm{B}T^*/(S_2 - S_1)$ and $-2k\mathrm{B}T^*/(S_2 - S_1)$. So, the temperature interval of phase co-existence is:

$$\Delta T \approx \frac{2k_\mathrm{B}T^*}{S_2 - S_1} - \left(\frac{-2k_\mathrm{B}T^*}{S_2 - S_1}\right) = \frac{4k_\mathrm{B}T^*}{S_2 - S_1} = \frac{4k_\mathrm{B}(T^*)^2}{E_2 - E_1} \tag{8.12}$$

Let us consider an instructive numerical example. When $T^* \sim 300$ K (ie, $k_\mathrm{B}T^* \sim 0.6$ kcal mol^{-1}) and $E_2 - E_1$ amounts to about 50 kcal mol$^{-1} \approx 100k_\mathrm{B}T^*$, which is typical of protein melting (ie, $E_2 - E_1 \approx 50/(0.6 \times 10^{23}) \sim 10^{-22}$ kcal per protein particle), then ΔT is about 10°. This means that molten and intact protein molecules co-exist in the range of a few degrees around the melting mid-point. However, when $E_2 - E_1$ is about 50 kcal per system (equivalent to melting a piece of ice as big as a bottle), then the co-existence range, ΔT, is about 10^{-23} degrees only.

In other words, "all-or-none" phase transitions of small systems are characterized by an energy jump within some temperature range. This is true to a much greater extent for first-order phase transitions in macroscopic systems. In these systems the range of the jump is almost infinitely narrow, while for macromolecules it covers several degrees (and thus still remains narrow as compared with the usual "observation range" from 0 to 100 °C). And in small oligopeptides no jump is observed: here the range of energy change can cover the entire "experimental window" of the temperatures studied.

A few words should be added concerning *second-order* phase transitions. Whereas first-order phase transitions are characterized by a *jump in the energy* of the system (along with its entropy, volume and density), the typical feature of

a second-order phase transitions is *an abrupt change* in the slope of the function $E(T)$, that is *a jump in the heat capacity*.

It should be stressed that a second-order transition occurs as the kink (for macro-systems) or the abrupt change (for macro-systems) in the $E(T)$ curve (Fig. 8.7) which occurs *not* in the middle of the subsequent more-or-less S-shaped dependence of energy (or other observed parameter) on the temperature.

Such a transition is typical, for example, when the system changes up to a certain temperature, and then this change stops (eg, in ferromagnetism, at a temperature below the Curie point, the heat is used to destroy the spontaneously magnetized state; and at a temperature above the Curie point, with the spontaneous magnetization already removed, the added heat is spent to increase the fluctuations only).

Microscopic analogs of the second-order phase transition have been found in proteins only recently, while the gradual and "all-or-none" transitions in polypeptides and proteins are known already for a long time. We will discuss them in detail later.

Figs. 8.5–8.7 show (deliberately) the cases when the system first changes slowly with increasing temperature, then rapidly, and then slowly again. These are selected to show that such "S-shaped" behavior of $E(T)$ is compatible with both a first-order phase transition and a gradual change of the system in the absence of any phase transition at all, as well as with a gradual change triggered by a second-order phase transition.

I ask you to pay attention to these examples because for unknown reasons, some non-physicists are prejudiced that a "transition" always corresponds exactly in the middle of any S-shaped profile, and that any "transition" that is not first-order is second-order. This is not true.

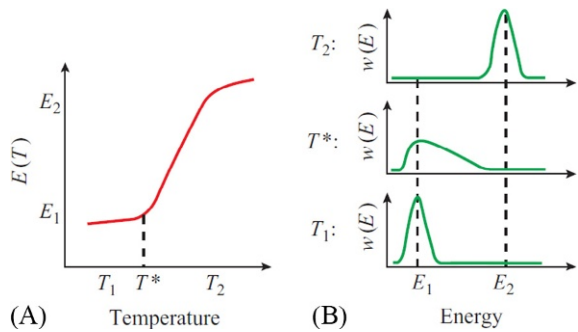


FIG. 8.7 Typical appearance of (A) the energy, $E(T)$, and (B) the energy distribution function, $w(E)$, for a second-order phase transition. T^* is the temperature of this phase transition. At this temperature, the distribution of energies, $w(E)$, becomes dramatically expanded, and the energy *begins* (or *stops*, depending on the direction of temperature change) its rapid change. Note that the second-order phase transition temperature T^* coincides with the *beginning* of this rapid energy change (and not with the middle, which is typical for "all-or-none" transitions shown in Fig. 8.6).

Note that although the curves $E(T)$ in Figs. 8.5–8.7 are alike, the behavior of the distribution curves, $w(E)$, for first-order (all-or-none) transitions is *utterly different* from the others. It is *only* "all-or-none" transitions that are characterized by two peaks on $w(E)$ curves reflecting the *co-existence* of two phases. Therefore, to see whether the transition between two extreme states is jump-like or gradual, it is *insufficient* to register a rapid change of energy or any other parameter in a narrow temperature range.

Some additional measurements are required, which will be the subject of discussion in a future lecture.

Now let us talk about the *kinetics of conformational changes*, or rather, about why their rate is sometimes extremely slow. What is "slow" here? Suppose you know the rate of one elementary step of the process. For example, it takes a residue ~1 ns to join the secondary structure. Also, you know that the chain contains 1000 residues. But the entire process takes not 1000 ns but 1000 s. That's what is "slow": orders of magnitude slower than expected either from the rate of the steps and their necessary number, or from the rate of diffusion. We have to understand the origin of this difference.

As mentioned, in some processes, the slow rate is caused by slow diffusion (which will be discussed soon) at high viscosity of the media, or by a necessity to make many steps in the course of reaction. However, the slow rate often—not always but often—results from the necessity of overcoming a high *free-energy barrier*. This is a very characteristic feature of "all-or- none" transitions, where the free-energy barrier separates two phases (Fig. 8.6); its value is always considerable here.

This barrier is very similar to the activation energy barrier of chemical reactions, although here it has both energy and entropy constituents (if the latter constituent is the main one, such a barrier is often called a "gate" (Frauenfelder, 2010)); later, we will see that the free energy barrier between the folded and unfolded states of a protein looks like a gate from the side of the unfolded state and as an energy barrier from the side of the folded state.

Let me remind you how to estimate the rate of such "barrier-overcoming" reactions using the classical theory of transition states (Emanuel and Knorre, 1984; Evans and Polanyi, 1935; Eyring, 1935; Pauling, 1970; Pelzer and Wigner, 1932).

First, let us consider the simplest process of a system transition from state 0 to state 1. On the pathway $0 \rightarrow 1$ let there be one barrier # (Fig. 8.8) and no "traps" (Fig. 8.9; the presence of "traps" complicates the process, but its nature remains unchanged).

Note: Kinetics of transition observed for each individual system (in, eg, "single-molecule experiments") looks *very* different from the "averaged" kinetics observed for a large ensemble of systems (molecules).

A single molecule does not "move" along the reaction coordinate (as it may seem from kinetics observed for a large ensemble, see the bottom panel in Fig. 8.8B): the ensemble demonstrates a smooth kinetics, while each single

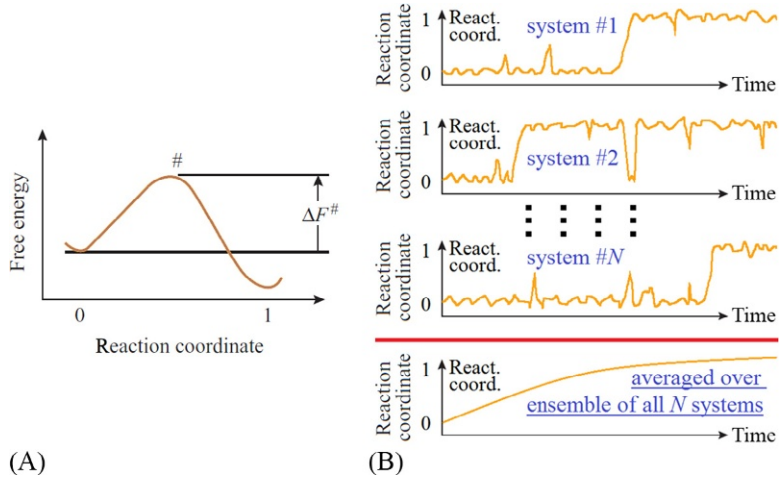


FIG. 8.8 (A) The free-energy barrier (activation barrier) # in the transition from the state "0" to state "1". $\Delta F^{\#}$ is the free energy of the barrier (ie, of the "transition state") counted off the preceding stable state 0. (B) Individual trajectories of overcoming the free energy barrier by single molecules (yellow, red, blue) and the overall kinetics of this process for a multitude of molecules.

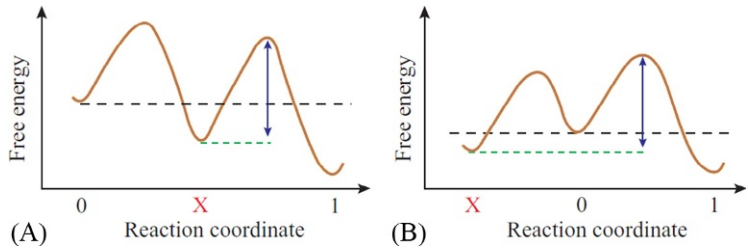


FIG. 8.9 A "trap": this may be either the on-pathway intermediate "X" (A) or an out-of-pathway (for the reaction $0 \rightarrow 1$) state "X" (B). The kinetic trap occurs provided it is more stable than the initial state "0" but less stable than the final state "1", and there is a higher free energy barrier (shown with an arrow) between "X" and "1" than between "0" and "X".

molecule tries to jump—falls back—tries to jump again—fails again—……—(see upper panels in Fig. 8.8B) and only finally succeeds.

If $\Delta F^{\#}$ is the free energy of the barrier relative to the free energy of the initial state, with $\Delta F^{\#} \gg k_B T$, if there are no "traps" (Fig. 8.9), and if n "particles" (molecules, systems, etc.) are in the initial state, then due to fluctuations, $n^{\#} \approx n \exp(-\Delta F^{\#}/k_B T)$ particles are on the barrier. Let it take each of these the time τ to jump from the barrier (τ being the time of "an elementary reaction step"). Then, during a time interval of about τ, all $n^{\#}$ "on-barrier" particles will cross the barrier. For all of the n particles present in state "0", the time of

performing $n/n^{\#}$ elementary steps is required to come to state "1", and this *time* of $0 \rightarrow 1$ transition amounts to:

$$t_{0\rightarrow 1} \approx \tau\left(n/n^{\#}\right) = \tau \exp\left(+\Delta F^{\#}/k_{\mathrm{B}}T\right) \tag{8.13}$$

The reciprocal

$$k_{0\rightarrow 1} \equiv 1/t_{0\rightarrow 1} \approx k_0 \exp\left(-\Delta F^{\#}/k_{\mathrm{B}}T\right) \tag{8.14}$$

is the *rate* of transition from 0 to 1; $k_0 \equiv 1/\tau$ is the rate of an elementary reaction step.

The rate of the reverse $1 \rightarrow 0$ transition amounts to:

$$k_{1\rightarrow 0} \approx k_0 \exp\left(-F^{\#}/k_{\mathrm{B}}T\right) \times \exp[-(F_0 - F_1)/k_{\mathrm{B}}T] \tag{8.15}$$

where F_0, F_1 are the free energies of states "0" and "1", respectively. To obtain this equation, we used the well-known ratio

$$k_{1\rightarrow 0}/k_{0\rightarrow 1} = \exp[-(F_0 - F_1)/k_{\mathrm{B}}T] \tag{8.16}$$

which follows from the fact that the equilibrium population of these two ("1" and "0") states, n_1^0 and n_0^0, must satisfy both the kinetic equation $n_1^0 k_{1\rightarrow 0} = n_0^0 k_{0\rightarrow 1}$ (accounting for zero flow of particles from one state to the other in equilibrium) and the thermodynamic ratio $n_1^0/n_0^0 = \exp[-F_1/k_B T]/\exp[-F_0/k_B T]$ (connecting the equilibrium probability of the particle being in either state with its free energy).

With the "trap" X present (Fig. 8.9), the $0 \rightarrow \mathrm{X}$ and $\mathrm{X} \rightarrow 1$ transition time is estimated in the same way. Then the total time of transition from 0 to 1 is the sum of the times of $0 \rightarrow \mathrm{X}$ and $\mathrm{X} \rightarrow 1$ transitions.

Note two further points:

First, if there are several transition pathways used *in parallel* (Fig. 8.10A), their *transition rates* must be *summed*:

$$k_{1+2+\ldots} = k'_{0\rightarrow 1^{\#}\rightarrow 1} + k''_{0\rightarrow 2^{\#}\rightarrow 1} + \ldots \tag{8.17}$$

Here $k'_{0\rightarrow 1^{\#}\rightarrow 1} \equiv 1/t'_{0\rightarrow 1^{\#}\rightarrow 1} = (1/\tau)\exp\left(-\Delta F^{\#1}/k_{\mathrm{B}}T\right)$ is the transition rate for the first pathway, $k''_{0\rightarrow 2^{\#}\rightarrow 1}$ is that for the second pathway, etc. So, if the parallel

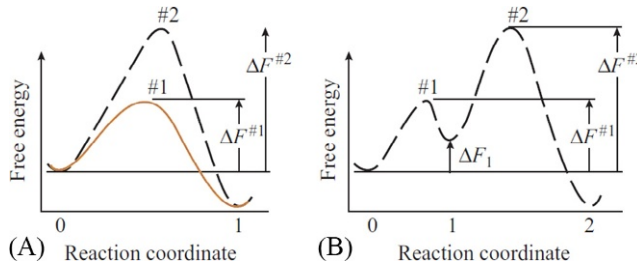


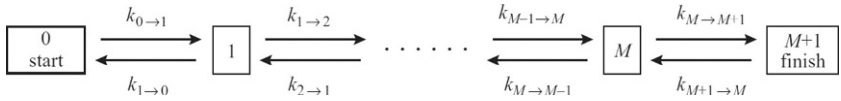
FIG. 8.10 The free energy barriers in parallel (A) and consecutive (B) processes.

pathways are not too numerous, the process time is determined by the most rapid of them, that is, by the one overcoming the *lowest* barrier.

Second, if there are several *consecutive* barriers on the transition path, then the individual barrier-overcoming *times* must be *summed*.

This statement is evident for the process shown in Fig. 8.9A, where the stable intermediate X is first accumulated before giving rise to the final state 1. However, it is far less evident for the process shown in Fig. 8.10B where the intermediate "1" is *un*stable. Moreover, in this case it should be specified that here "the individual barrier-overcoming time" implies the time required for a particle to reach the on-barrier position from the *deepest* prior minimum of the free energy, and *not* from the *immediately* preceding one (eg, in Fig. 8.10B, the height of barrier #2 must be taken relative to state 0 not to state 1).

To prove this (I will provide the idea only without boring you with calculations), let us consider the process



(where $k_{i\to i+1}$ is the rate of transition from i to $i+1$, and $k_{i+1\to i}$ is the rate of inverse transition from $i+1$ to i).

A general solution of this differential kinetic equation pertaining to this problem has been given by a Moscow scientist, Rakowski (1907), but now we are interested only in a simple estimate (see Becker and Döring, 1935) of the rate of this process for a case where the free energies of all intermediate states (1, 2,…, M) are higher by many times k_BT than the free energies of both initial (0) and final ($M+1$) states.

Because of their high free energies, all intermediate states accumulate only a few molecules (as compared with the total number of the molecules in the initial and final states). Therefore, the rate at which the numbers of molecules change is also very low for the intermediate states as compared with that for the initial and final states. In other words, it can be assumed that the flow rate is approximately constant over the entire reaction pathway. Thus:

$$\begin{aligned} -\mathrm{d}n_0/\mathrm{d}t &= k_{0\to1}n_0 - k_{1\to0}n_1 = k_{1\to2}n_1 - k_{2\to1}n_2 = \ldots . \\ &= k_{M\to M+1}n_M - k_{M+1\to M}n_{M+1} \equiv \mathrm{d}n_{M+1}/\mathrm{d}t \end{aligned} \tag{8.18}$$

(This assumption is called the "*quasi-stationary*," or "*steady-state*" *approximation*, and is widely used in chemical kinetics; Emanuel and Knorre, 1984; Pauling, 1970). Solution to the above equations gives the result:

$$\begin{aligned} t_{0\to\cdots\to M+1} = {} & (1/k_{0\to1}) + (1/k_{1\to2})\exp(+F_1/k_BT) \\ & + \cdots + (1/k_{M\to M+1})\exp(+F_M/k_BT) \end{aligned} \tag{8.19}$$

where $1/k_{i\to i+1}$ is the barrier i overcoming time, provided the starting point was the immediately preceding state i, and ΔF_i is the free energy of the intermediate state i relative to the free energy of the initial state. To obtain this equation we used Eq. (8.16).

One last point: since in accordance with Eq. (8.13) $1/k_{i-1\to i}=\tau_i \exp[(\Delta F^{\#i}-\Delta F_{i-1})/k_BT]$, then $(1/k_{i-1\to i})\exp(\Delta F_{i-1}/k_BT)=\tau_i \exp(\Delta F^{\#i}/k_BT)$. This is just the time required to overcome the individual barrier $\#i$, if it were the only one on the pathway of the process $0\to M+1$; let us refer to this time as $t_{0\to\#i\to}$. Then

$$t_{0\to\cdots\to M+1}=t_{0\to\#1\to}+t_{0\to\#2\to}+\cdots+t_{0\to\#M+1} \tag{8.20}$$

which proves the statement that the time of a *consecutive* reaction is the *sum of the times* required to overcome individual barriers the heights of which are taken relative to *the deepest* of the prior free energy minima (Eq. (8.20) gives an approximate solution; a more accurate solution based on generalization of the steady-state approach one can find in Finkelstein, 2015).

Incidentally, Eq. (8.20) shows that if the sequential barriers are not too numerous, the time of the process is determined simply by the *highest* one among them.

Up to now, we paid attention to the barriers and left aside $k_0\equiv 1/\tau$, the rate of an elementary step. The value k_0 is often (Pauling, 1970) taken as $k_{0,el}=k_BT/h$ (where h is the Planck's constant); k_BT/h is about $10^{13}\ s^{-1}$ at $T\approx 300$ K: this is the frequency of attacks on the barrier under the action of thermal vibrations.

Considering a multistep process, it is often convenient to express its rate through the rates of one-step transitions between intermediate free energy minima and the heights of these minima (Fig. 8.11). To this end, we can:

1. Take the rate of transition between two adjacent minima as the rate of an elementary step:

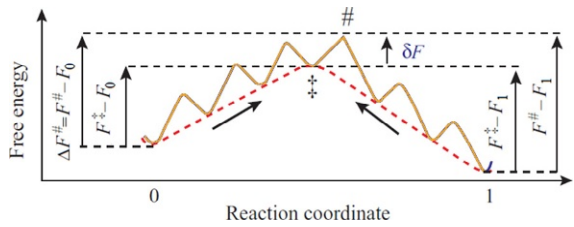


FIG. 8.11 Overcoming of the free energy barrier on the pathway from the stable state "0" to the stable state "1" (the left arrow), and from "1" to "0" (the right arrow). F_0, F_1 are the free energies of the stable states "0", "1"; $F^{\ddagger}$ is that of the most unstable intermediate metastable state ‡, and $F^{\#}$ is that of the transition state (ie, the most unstable state on the reaction pathway). δF is the height of the free energy barrier for one step.

$$k_{0,\text{step}} = k_{0,\text{el}} \exp(-\delta F / k_B T) \tag{8.21}$$

(where δF is the free energy barrier that is to be overcome during one step, see Fig. 8.11)

2. Take the maximum free energy of an intermediate metastable state, $F^{\ddagger}$, as the effective barrier free energy for the whole process; then:

$$\begin{aligned} k_{0\to 1} &= k_{0,\text{el}} \exp\left[-\left(F^{\#} - F_0\right)/k_B T\right] = k_{0,\text{el}} \exp\left[-\left(F^{\ddagger} + \delta F - F_0\right)/k_B T\right] \\ &= k_{0,\text{step}} \exp\left[-\left(F^{\ddagger} - F_0\right)/k_B T\right] \end{aligned} \tag{8.22}$$

This equation looks like Eq. (8.14) with $k_{0,\text{step}}$ in place of k_0 and $F^{\ddagger} - F_0$ in place of $\Delta F^{\#} \equiv F^{\#} - F_0$.

The same substitutions and explanations are, of course, equally true for $k_{1\to 0}$ and $t_{1\to 0}$.

Finally, let us consider the diffusion rates. As I have already said, the existence of a free energy barrier is suggested by a reaction rate that is much lower than the rate of diffusion. And what is the typical diffusion time? To have some idea, let us talk a little about diffusion.

Before we do this, it is useful to estimate how long a molecule needs to forget the direction of its movement and to start diffusing. That is, we have to know how long it takes for the molecule's kinetic energy to dissipate because of friction against a viscous fluid. One can show that this occurs in picoseconds (ps).

Indeed, the particle's movement in a viscous fluid is described by the Newton equation $m(\mathrm{d}v/\mathrm{d}t) = F_{\text{frict}}$ where m is the particle's mass, $\mathrm{d}v/\mathrm{d}t$ is acceleration and F_{frict} is the force of friction. The mass can be estimated as $m = \rho V$, where ρ is the particle's density and V its volume. The friction force, for a spherical particle, is $F_{\text{frict}} = -3\pi D\eta v$ according to Stokes' law, where D and v are the spherical particle's diameter and speed and η is the viscosity of the fluid. The equation:

$$m(\mathrm{d}v/\mathrm{d}t) = -3\pi D\eta v \tag{8.23}$$

determines the time

$$t_{\text{kinet}} \approx m/(3\pi D\eta) \tag{8.24}$$

typical of the friction-caused movement damping. In fact, $t_{\text{kinet}} \approx 0.1\rho D^2/\eta$, since $m \approx \rho D^3$, and $3\pi \approx 10$. Since $\rho \approx 1$ g cm^{-3} for all the molecules we deal with, and $\eta \approx 0.01$ g cm^{-1} s^{-1} for water (see any databook), we have:

$$t_{\text{kinet}} \approx 10^{-13}\text{s}(D/\text{nm})^2 \tag{8.25}$$

where (D/nm) is the particle's diameter expressed in nanometres. If the molecule is not a ball but an ellipsoid with axes d_1, d_2, d_3, then the effective $D^2 = (d_1 d_2 + d_2 d_3 + d_3 d_1)/3$ (Landau and Lifshitz, 1987).

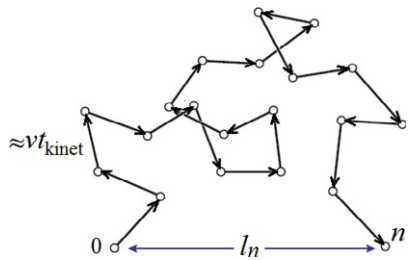
This means that the kinetic energy of a small ($D \approx 0.3$ nm) molecule (eg, water) dissipates within $\sim 10^{-14}$ s, of a small protein ($D \approx 3$ nm) within $\sim 10^{-12}$ s, and of a large protein ($D \approx 10$ nm) within $\sim 10^{-11}$ s. Thus, for aqueous solutions, the typical time is a picosecond.

Note: For a more viscous environment, a membrane for example, the kinetic energy dissipation is proportionally faster.

Of course, collisions with other molecules compensate for this energy loss. But the direction of the initial movement is forgotten within a picosecond.

Now we can turn to the diffusion movement of a molecule.

The heat-maintained kinetic energy of each particle, $mv^2/2$, amounts, on average, to about $k_B T$. The particle "memorizes" the direction of its movement for a time t_{kinet}. During this time, the time of one step, it covers a distance $\Delta l \approx v\, t_{kinet}$ (this distance is called the mean free path of the molecule; in water, it can be estimated as ≈ 0.13 nm $(D/\text{nm})^{1/2}$; see Problem 8.8). Then the direction of its movement changes, and it covers approximately the same distance Δl in some new direction. That is, at each step its displacement is about $\pm\Delta l$. The mean square displacement of the molecule from the initial point grows proportionally with time. Indeed, if, after n steps, the particle is moved by a distance l_n in some direction, then its displacement after $n+1$ steps is $l_n + 1 = l_n \pm \Delta l$, and $(l_n+1)^2 = (l_n \pm \Delta l)^2 = l_n^2 \pm 2 l_n \Delta l + \Delta l^2$. That is, since the mean value of the term $\pm 2 l_n \Delta l$ is zero, $(l_n+1)^2 = l_n^2 + \Delta l^2$ on average (recollect coil where the squared end-to-end distance is proportional to the chain length and see the following scheme as an illustration):



The particle makes t/t_{kinet} steps within time t; thus, its mean square displacement after time t is

$$l_t^2 = (t/t_{kinet})\, \Delta l^2 \tag{8.26}$$

Since $\Delta l \sim v\, t_{kinet}$,

$$l_t^2 \sim (t/t_{kinet})(v t_{kinet})^2 = t\left(v^2 t_{kinet}\right) \tag{8.27}$$

and since $t_{kinet} \approx m/(3\pi D\eta)$, and $mv^2/2 \approx k_B T$,

$$l_t^2 \sim t[k_B T/(1.5\pi D\eta)] \tag{8.28}$$

This answer is only approximate, since we have used the symbols "≈" (approximately equal) and "∼" (equal in the order of magnitude) many times. However, as usually happens in such cases, the approximate answer is close to the precise answer (which requires much more refined calculations). The precise answer is:

$$l_t^2 = t[2k_B T/(\pi D\eta)] \equiv t[6D_{\text{diff}}] \tag{8.29}$$

Here $D_{\text{diff}} = k_B T/(3\pi D\eta)$ is the Einstein (more accurately: Stokes–Einstein, 1905—Sutherland, 1905—Smoluchowski, 1906) value of the diffusion coefficient for a ball of diameter D in the medium with viscosity η at temperature T (Landau and Lifshitz, 1987).

The characteristic diffusion time is the time spent by a molecule in diffusing for a distance of its diameter D. It is easy to estimate this time from Eq. (8.29):

$$t_{\text{diff}} = \left(\pi D^3 \eta\right)/(2k_B T). \tag{8.30}$$

Since water viscosity $\eta \approx 0.01$ g $(\text{cm}^{-1}\,\text{s}^{-1})$, and $k_B T \approx 600$ cal $\text{mol}^{-1} \approx$ 2500 J mol^{-1} at room temperature,

$$t_{\text{diff}} \approx 0.4 \times 10^{-9} s(D/nm)^3, \tag{8.31}$$

where (D/nm) is again the particle's diameter expressed in nanometres.

It is possible to show that a particle's inversion takes approximately the same time (the inversion can be regarded as displacement of the particle's pole by a distance of $\sim D$).

It is useful also to bear in mind that diffusion at the 1 nm distance takes $\approx 0.4 \times 10^{-9}$ s (D/nm).

Thus, we can conclude that in water, the typical diffusion time of a molecule falls within a nanosecond range: within this time a molecule driven by collisions with its fellow molecules covers a distance equal to its size and/or is overturned.

Note: The above estimates are for the aqueous environment. For a more viscous environment (eg, in a cell where viscosity is higher by an order of magnitude, according to some estimates), the times are proportionally longer.

Any process that takes much more time than diffusion allows us to suggest the existence of a free energy barrier on its pathway. For inter-molecular reactions (or reactions between remote chain regions), this barrier is created, in part, by the entropy loss required to bring together the reacting pieces.

REFERENCES

Becker, R., Döring, W., 1935. Kinetische Behandlung der Keimbildung in Übersättigten Dämpfen. Annalen der Physik (in German) 24, 719–749.

Einstein, A., 1905. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Annalen der Physik (in German) 322, 549–560.

Emanuel, N.M., Knorre, D.G., 1984. The Course in Chemical Kinetics, fourth Russian ed. Vysshaja Shkola, Moscow. Chapters III (§ 2), V (§§ 2, 5).

Evans, M.G., Polanyi, M., 1935. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. Trans. Faraday Soc. 31, 875–894.

Eyring, H., 1935. The activated complex in chemical reactions. J. Chem. Phys. 3, 107–115.

Feynman, R., Leighton, R., Sands, M., 1963. The Feynman Lectures on Physics, vol. 1. Addison-Wesley Publishing Company, Inc., Reading, MA (Chapter 46).

Finkelstein, A.V., 2015. Time to overcome the high, long and bumpy free-energy barrier in a multi-stage process: the generalized steady-state approach. J. Phys. Chem. B 119, 158–163.

Frauenfelder, H., 2010. The Physics of Proteins. An Introduction to Biological Physics and Molecular Biophysics. Springer, NY (Chapter 13).

Landau, L.D., Lifshitz, E.M., 1987. Fluid Mechanics (Volume 6 of A Course of Theoretical Physics), second ed. Butterworth-Heinemann, Oxford. §§ 58–60.

Landau, L.D., Lifshitz, E.M., 1980. Statistical Physics (Volume 5 of A Course of Theoretical Physics), third ed. Elsevier, Amsterdam. §§ 7–9, 14–15, 28, 31, 137–138, 150.

Pauling, L., 1970. General Chemistry. W.H. Freeman & Co., NY (Chapter 16).

Pelzer, H., Wigner, E., 1932. Über die Geschwindigkeitskonstante von Austauschreaktionen. Z. Phys. Chem. (in German) B15, 445–471.

Rakowski, A., 1907. Kinetik der Folgereaktionen erster Ordnung. Z. Phys. Chem. (in German) 57, 321–340.

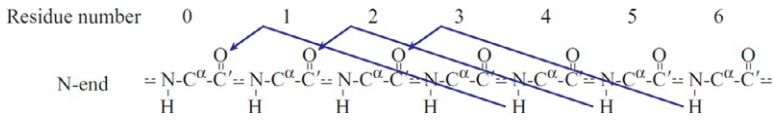
Sutherland, W., 1905. Dynamical theory of diffusion for non-electrolytes and the molecular mass of albumin. Phil. Mag. 9, 781–785.

von Smoluchowski, M., 1906. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. Annalen der Physik (in German) 326, 756–780.

Lecture 9

With basic physics learned, let us move on to discuss the stability of the secondary structure and the kinetics of its formation. We will now consider only *homo*-polypeptides, ie, chains formed by identical amino acid residues.

We start by considering an α-helix. The conformations of its first three residues (1, 2, 3) in this helix are fixed with its first hydrogen bond $(CO)_0 \leftarrow (HN)_4$; the next hydrogen bond, $(CO)_1 \leftarrow (HN)_5$, additionally stabilizes the conformation of only one residue (residue 4); the hydrogen bond $(CO)_2 \leftarrow (HN)_6$ provides additional binding for residue 5, and so on.



Thus, n residues are fixed by $n-2$ hydrogen bonds. Let us consider the free energy of formation of such a helix from a coil in aqueous solution (a coil is a polymer without any fixed structure and without interactions between non-neighboring residues). This free energy is given by:

$$\Delta F_\alpha = F_\alpha - F_{\text{coil}} = (n-2)f_H - nTS_\alpha = -2f_H + n(f_H - TS_\alpha) \qquad (9.1)$$

Here f_H is the free energy of formation of a hydrogen bond in the α-helix. Apart from the free energy of the H-bond per se (which, as you remember, is not just the energy as would be the case in a vacuum, but includes both the energy and entropy of the subsequent H-bond rearrangement in the aqueous environment), it also includes the free energy of other interactions accompanying formation of the H-bond in the helix. S_α is the entropy loss caused by fixation of one residue in the helix.

As you see, $\triangle F_\alpha$ has two terms. One of them $(-2f_H)$ is independent of the helix length; the quantity

$$f_{\text{INIT}} = -2f_H \qquad (9.2)$$

is known as the free energy of helix initiation (actually, f_{INIT} reflects both helix initiation and termination). The other term, $n(f_H - TS_\alpha)$, is directly proportional to the helix length; the quantity

$$f_{\text{EL}} = (f_H - TS_\alpha) \qquad (9.3)$$

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00009-0

is known as the free energy of helix elongation per residue. Generally, we have:

$$\Delta F_{\alpha} = f_{\mathrm{INIT}} + n f_{\mathrm{EL}} \tag{9.4}$$

The relationship between the probabilities of the purely helical state of an n-residue-long sequence and its purely coil (free of any helical admixtures) state is expressed as:

$$\exp\left(-F_{\alpha}/kT\right) = \exp\left(-f_{\mathrm{INIT}}/kT\right) \times \left[\exp\left(-f_{\mathrm{EL}}/kT\right)\right]^{n} = \sigma s^{n} \tag{9.5}$$

Here, I have used the conventional notation (Schulz and Schirmer, 1979/2013): $s = \exp(-f_{\mathrm{EL}}/kT)$, the helix elongation parameter; $\sigma = \exp(-f_{\mathrm{INIT}}/kT)$, the helix initiation parameter.

It is obvious that $\sigma \ll 1$, since $\sigma = \exp(-f_{\mathrm{INIT}}/kT) = \exp(+2f_{\mathrm{H}}/kT)$; and the free energy of a hydrogen bond is a large negative value of about several kT.

The quantity $\exp(-\triangle F_{\alpha}/kT) = \sigma s^{n}$ is simply the *equilibrium constant* for the two states ("α" and "coil") of an n-residue-long sequence.

Prior to discussing the ways of experimentally determining the σ and s values, let us see whether under varying conditions (temperature, solvent, etc.) the helix forms gradually or through an "all-or-none" transition.

On the face of it, such a distinct structure as the α-helix should be "frozen out" of the coil by a phase (ie, "all-or-none") transition, like ice out of water.

However, *Landau's theorem* states that first-order phase transitions never occur in a system consisting of two *one-dimensional* (1D) phases (Landau and Lifshitz, 1980). Let me try to explain this.

First of all, what does one-dimensionality mean? It means that the size (and hence, the free energy) of the phase interface is independent of the phase sizes. In these terms, both helical and coil conformations of a polypeptide are 1D. Fig. 9.1 shows that the interface between the helix and the coil regions is independent of their lengths, unlike the interface of the 3D phases (eg, of a piece of ice with surrounding water). Consequently, the free energy on the helix boundaries does not depend on the helix size, while the free energy of a 3D phase (ice) increases as $n^{2/3}$ with increasing number n of the particles involved in this phase.

Now what does "forming through a first-order phase transition" mean? This means that at a transition temperature either phase can be stable, but their mixture (eg, the mixture of ice and water) is unstable owing to increasing free energy. You must not be misled by the picture of a floating in water piece of

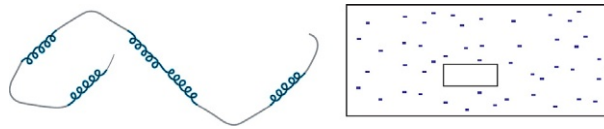


FIG. 9.1 Comparison of 1D (coil with helices) and 3D (a piece of ice in water) systems. The size of the interface between the helix and the coil is independent of their lengths, while the interface of the 3D piece of ice with water varies with its size.

ice: this state is unstable at any temperature (owing to the additional free energy at the interface between ice and water), and with time, at a fixed temperature, ice will either melt or take up the entire water, provided there is no flow of underground heat, streams on other interfering non-equilibrium factors.

Is phase co-existence in a 3D system favorable? *No, it is not*. Why?

Let us return to Fig. 9.1 and consider the temperature at which infinite water and infinite ice have equal values of the free energy (that is a condition of the "mid-transition"). If the floating piece of ice consists of n molecules, the interface free energy is proportional to $\xi n^{2/3}$, where $\sim 6n^{2/3}$ is the characteristic number of interface molecules (suggesting that ice has a more or less cubic form), and $\xi/6 > 0$ is the interface free energy of each of them. (Note that if $\xi < 0$, the thermodynamically favorable (in this case) "mixing up" occurs on the molecular scale, and the two phases do not emerge at all.) Consequently, the ice surface increases the free energy by $\xi n^{2/3}$. True, the piece of ice also possesses positional entropy, since its position in the vessel can vary. But this entropy never exceeds the value of the order of $k \ln(N)$, if there are N molecules (ie, N ice initiation points) in this vessel. In total, the free energy of this piece of ice amounts to about $[\xi n^{2/3} - kT \ln(N)]$. However, at large N values, the logarithm grows only slightly. If the piece of ice occupies a considerable part of the vessel (eg, $n \sim N/10$) and N is very large (eg, 10,000,000,000), then $\ln(N)$ (here, 23) is very small compared with $(N/10)^{2/3}$ (here, 1,000,000); in other words, the interface term $\xi n^{2/3}$ predominates, and this term is unfavorable for the formation of a piece of ice. Therefore, in a 3D system, macroscopic phases fall apart thereby making possible a first-order transition. (Unstable ice pieces of only a few molecules can be neglected as they are no more than microscopic, local fluctuations in water.)

And is phase co-existence in a 1D system favorable? *The answer appears to be yes*.

Let us return to the "mid-transition temperature" at which the helix and the coil have the same free energy, ie, $f_{\mathrm{EL}} = 0$. The free energy of the helix boundaries, f_{INIT}, is independent of both helix and coil lengths. The positional entropy of an n-residue-long helix in an N-residue-long chain is $k \ln(N - n)$. In total, the free energy of the helix floating in the sequence is $f_{\mathrm{INIT}} - k \ln(N - n)$. At large values of N and not too large n (eg, $n \sim N/10$ or $\sim N/2$), the term containing $\ln(N - n)$ *always* predominates over the constant (f_{INIT}); this logarithmic term reduces the free energy and promotes insertion of the helix into the coil (as well as insertion of the coil into the helix). That is why, in a 1D system, phase division *does not* happen; the phases tend to mix up, and therefore, a first-order transition (ie, the "all-or-none" type transition) becomes impossible, provided the sequence is sufficiently long. Thus, the Landau theorem is proved.

Note: Strictly speaking, unlike α-helix melting, that of the DNA double helix does not come within the Landau theorem, since the double-stranded DNA is *not* a 1D system: the DNA molten region is a spatial loop closed by double helices at its ends. The loop closing provides an additional contribution to the free

energy of the loop boundaries, which, according to Flory's formula for loops (Flory, 1969), grows logarithmically with increasing loop length.

Now we come to the question, "At which characteristic chain lengths do the coil and helical phases begin to mix up?" Or rather, "What characteristic length n_0 of the helical segment corresponds to the midpoint of the helix-coil transition?"

Let us consider an N-residue sequence at mid-transition temperature when the values of the free energies of the helix and coil are equal, ie, $f_{EL}=0$. Then the free energy of helix elongation (and coil elongation as well) is zero, that of helix initiation is f_{INIT}, the number of possible positions of the ends of a helix in the n-residue chain region is about $n^2/2$ (the helix can be started and ended anywhere; the condition that the helix must contain ≥ 3 residues is not significant when $n \gg 3$). The free energy of the helix is unaffected either by the helix position or by its length. To obtain a qualitative estimate, minor things (numerals in equations) can be neglected, and only major ones (letters in equations) must be taken into account. Then the entropy of the ends is $\approx k \times 2\ln(n)$; and the total free energy of insertion of a portion of the new phase (a helix with fluctuating ends into the n-residue coil or a coil into the n-residue helix) is $\approx f_{INIT} - 2kT\ln(n)$. If this free energy is greater than zero, the insertion of the new phase will not happen; if it is less than zero, the insertion will happen and may be repetitive until the average phase length n exceeds n_0 that can be found from the equation $f_{INIT} - 2kT\ln(n_0)=0$. Thus, at the midpoint of the helix-coil transition the characteristic lengths of their fragments is (see Zimm and Bragg, 1959; Schulz and Schirmer, 1979/2013).

$$n_0 \approx \exp(+f_{INIT}/2kT) \equiv \sigma^{-1/2} \tag{9.6}$$

Experimentally, the mid-point of this transition is the point (temperature) corresponding to 50% helicity of a very long polypeptide (as mentioned earlier, the helicity of a polypeptide is usually measured using CD spectra; at 50% helicity, the polypeptide CD spectrum represents a half-sum of the spectra of the polypeptide coil conformation and its totally helical conformation). At this point $f_{EL}=0$, ie, $s=\exp(-f_{EL}/kT)=1$.

The n_0 value can be found as the sequence length that provides 12% helicity at $s=1$. (I will not prove this numerical estimation, as it is beyond the scope of these lectures. You can try and do it yourself using Appendix B.)

I would just like to explain why the helicity of an n_0-residue chain is several times lower than that of a very long sequence (ie, <50%). This is because this chain can be either completely in the coil state (in this case, its free energy is zero), or include a one helix/coil mixture (with an additional free energy of about $f_{INIT} - kT\ln[(n_0)^2/2] = +kT\ln 2 > 0$, ie, having a <50% probability), where the helix covers only some part of the chain.

Finally, with n_0 known, we can calculate f_{INIT} and σ. For most amino acids, $n_0 \approx 30, f_{INIT} \approx 4$ kcal mol^{-1} and $\sigma \approx 0.001$ (see Ptitsyn and Finkelstein, 1979a,b; Muñoz and Serrano, 1994, and references therein).

Now we can find the free energy of H-bonding (together with all the interactions accompanying formation of a hydrogen bond in an α-helix): according to Eq. (9.2), $f_H = -f_{INIT}/2 \approx -2$ kcal mol^{-1}. Also, it is possible to determine the loss of conformational entropy caused by fixation of a residue in the α-helix: according to Eq. (9.3), at $f_{EL} = 0$, $TS_\alpha = f_H \approx -2$ kcal mol^{-1}.

Both parameters of helix stability, f_{EL} and f_{INIT}, depend on conditions (eg, temperature), but what greatly influences the stability is the deviation of f_{EL} from 0. The reason is that this deviation is multiplied by the large value n_0 (for an n_0-residue helix), and only in this way does it appear in the free energy of the helix. When the quantity $f_{EL}n_0/kT$ is close to +1 (or more precisely: $f_{EL}n_0/kT = +2$), the helicity almost disappears, and with $f_{EL}n_0/kT \leq -2$ the coil almost ceases to exist. Consequently, in very long ($N \gg n_0$) polypeptide chains the helix-coil transition occurs in a region of $-2/n_0 < f_{EL}/kT < +2/n_0$, ie, in a region of $-0.07 < f_{EL}/kT < 0.07$ at $n_0 \approx 30$ (Fig. 9.2). This is an example of an abrupt, cooperative, but *not* phase transition (since its width does not tend to zero with increasing length of the chain).

The stability of the α-helix usually decreases with increasing temperature and added polar denaturants; and it increases with added weakly (less than water) polar solvents (which increase the price of H-bonds).

To measure the effect of amino acid residues on helix stability, short ($\sim n_0$ or less) polypeptides are currently most often used. They can house only one helix, and therefore, the effect of each amino acid replacement on the helicity can be estimated most easily. Now it is known that the contribution of an amino acid residue to the helix stability ranges from alanine, the most "helix-forming" residue, with $s \approx 2$, ie, $f_{EL} \approx -0.4$ kcal mol^{-1}, to glycine, the most (but for Pro) "helix-breaking" residue, with $f_{EL} \approx +1$ kcal mol^{-1}, ie, $s \approx 0.2$, while the majority of other residues have f_{EL} close to zero (Finkelstein et al., 1976; Muñoz and Serrano, 1994). Proline (an *imino* acid with no NH group to participate in the helix-forming H-bond) has a considerably lower value of s (0.01–0.001; it has not been accurately measured yet).

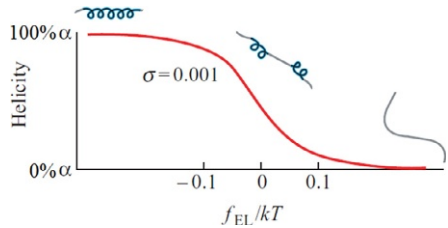


FIG. 9.2 A finite width is typical of any helix-coil transitions, even those in infinitely long chains. This is an example of a cooperative transition, which is *not* a phase transition: at a low value of the helix initiation parameter ($\sigma \ll 1$) it is caused by a certain small change (much less than kT) in the value of f_{EL}, which shows that a "transition unit" involves many chain residues but far from all of them.

Earlier, the estimates of this kind were made (in particular, by one of us, O.B.P.) using statistical co-polymers (eg, chains with a random mixture of 80% Glu and 20% Ala); it is in this way that the first—and hence the most important—estimates were obtained (see von Dreele et al., 1971; Platzer et al., 1972; Snell and Fasman, 1973; Ptitsyn and Finkelstein, 1979a,b, and references therein). But with the advent of pre-set sequence synthesis, the use of such random co-polymers became a thing of the past.

Also, potentiometric titration was used to measure the helical (as well as β-structural) state stability (the quantity f_{EL}) in polypeptides containing acidic or basic side chains (eg, in poly(Glu) or poly(Lys)) (Bychkova et al., 1971; Barskaya and Ptitsyn, 1971; Pederson et al., 1971; Mandel and Fasman, 1975; Walter and Fasman, 1977). The idea of this approach is that by charging a helix, we destroy it (because in the helix the side chain charges are closer to one another than in the coil, and, hence, their repulsion is stronger). So, the helix stability can be calculated from the dependence of the total charge and helicity of the chain on the medium's pH.

Unfortunately, consideration of this interesting method in more detail is beyond the scope of this course.

Using short peptides with pre-set sequences, an estimate can even be made as to how the helicity is affected by each single amino acid replacement (Padmanabhan et al., 1990; Fersht, 1999) at a given position in the peptide, ie, in fact, as dependent on the residue position about the N- and C-termini of the helix (and on the residues surrounding the residue in question). The side chains, and in particular charged ones, interact with these termini in opposite ways, because, as mentioned, the N-terminus of the helix houses the main-chain NH groups free of hydrogen bonds (and the resulting partial charge of the α-helix N-terminus is equal to $+e/2$), while its other terminus holds free CO groups (with the total partial charge $-e/2$, half the electron charge (Ptitsyn and Finkel'shtein, 1970; Finkelstein and Ptitsyn, 1976).

Similar approaches are used to measure the stability of the β-structure in polypeptides. However, they are less developed, since the β-structure aggregates strongly. Currently, β-structure stability is measured right in the proteins by estimating the effect on protein stability of replacements of each of its surface β-structural residues. It is shown that contribution of amino acid residues into β-structure stability ranges from ≈ -1 to $\approx +1$ kcal mol^{-1} (Finkelstein, 1995). The ability of various residues to stabilize α- and β-structures will be considered in Lecture 10.

Now let us consider *the rate of formation* of the secondary structure in peptides.

Experiment shows that α-helices are formed very rapidly: within ~0.1 μs a peptide of 20–30 residues adopts the helical conformation (Williams et al., 1996; Thompson et al., 1997); such rapid measurements require a pico- or nanosecond laser-induced temperature jump of the solution. Consequently, the rate of helix extension is at least a residue per several nanoseconds.

I said "at least" because the rate of helix formation depends not only on its extension rate but also on how rapidly the first "nuclei" of the helical structure appear. Initiation of the helix requires overcoming the activation barrier; therefore, the formation of the first turn is the slowest step, and subsequent growth of the helix is rapid.

Hence, it is possible that nearly the entire observation time might be taken by helix initiation, with its elongation being far more rapid. Let us consider this in more detail (see Galzitskaya et al., 2002, and references therein).

The typical dependence of the free energy of a helix on its length is illustrated in Fig. 9.3. Even if $f_{EL} < 0$, ie, when the rather long helix is stable, formation of its first turn requires overcoming an activation barrier as high as f_{INIT}.

Does the experimentally measured time of α-helix formation agree with our current understanding of this process?

According to the transition state theory, this first step of helix formation (formation of the first helix turn in the given place of the chain) takes the time

$$t_{INIT\_\alpha} = \tau \ \exp(+f_{INIT}/kT) \tag{9.7}$$

where τ is the time of an elementary step (here, helix elongation by one residue), and the exponent allows for low occupancy of the barrier state possessing (by definition of a barrier) the highest free energy. By definition, $\sigma = \exp(-f_{INIT}/kT)$; thus

$$t_{INIT\_\alpha} = \tau/\sigma \tag{9.8}$$

However, initiation can occur anywhere in the future helix, and its length (even in very long chains) is limited by $\sim n_0 = \sigma^{-1/2}$. Consequently, the typical time of initiation of the first turn *anywhere* in the future helix is n_0 times less, and $t_{INIT\_\alpha}/n_0 = \tau/(\sigma n_0) = \tau/\sigma^{1/2}$.

Propagation of the helix to all its $\sim n_0$ residues takes as much time as the initiation, $\sim \tau n_0 = \tau/\sigma^{1/2}$. Thus, half of the whole transition time, roughly, is spent on helix initiation anywhere in the sequence, and the rest on elongation.

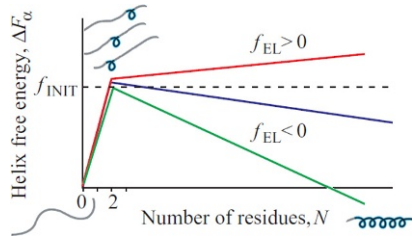


FIG. 9.3 Typical dependence of the α-helix free energy ($\triangle F\alpha$) on the number (N) of amino acid residues involved in the helix, with various free energies of helix elongation (f_{EL}). When $f_{EL} < 0$, a long helix is stable but its initiation requires overcoming an activation barrier as high as f_{INIT}. When $f_{EL} > 0$, a helix of any length is unstable, and therefore it cannot form. Note that the α-helix-initiating turn can be formed in any place of the future α-helix.

This gives the total helix-coil transition time as approximately $2\tau/\sigma^{1/2}$, and the half-time (ie, the characteristic time) of the whole transition is:

$$t_\alpha \sim \tau/\sigma^{1/2} = \tau \ \exp(+f_{\text{INIT}}/2kT) \tag{9.9}$$

The time of helix elongation (τ) has been estimated as about a residue per a few nanoseconds (Zana, 1975; Cummings and Eyring, 1975; Thompson et al., 1997), and f_{INIT} as about 4 kcal mol^{-1} (Barskaya and Ptitsyn, 1971; Ptitsyn and Finkelstein, 1979a,b). Thus, theoretically estimated t_α is about hundreds of nanoseconds, in accordance with experiment.

The kinetics of α-helix formation is relatively simple: all these are formed rapidly. The kinetics of β-structure formation is far more complex and interesting.

Experiment shows that the β-structure often forms extremely slowly in water-soluble polypeptides. It may take hours and even weeks, although sometimes the β-structure folds within milliseconds (Wooley and Holzwarth, 1970; Snell and Fasman, 1973; Fukada et al., 1989), and β-hairpins even faster (Muñoz et al., 1997, 1998). What is the reason for that? Surprisingly, the folding rate of proteins containing β-structures is not much lower than that of α-helical proteins (Fersht, 1999); we will discuss this later on. How do they manage? And what controls the anomalous rate of β-structure formation in large water-soluble polypeptides: slow initiation or slow elongation?

Let us start with formation of β-hairpins. Experiment shows that they are formed rather rapidly, though not as rapidly as α-helices: within ~5 μs a peptide of ~20 residues adopts the β-hairpin conformation (Muñoz et al., 1997, 1998). Does this agree with our current understanding of formation of β-hairpins?

Formation of a beta turn in the given place of the chain must take the time $\sim\tau_\beta \exp(+f_{\text{INIT\_}\beta}/kT)$, where the free energy $f_{\text{INIT\_}\beta}$ of beta turn formation must be similar to that of α-turn (f_{INIT}), and the time of β-structure elongation $\tau_\beta \approx \tau$. However, unlike in the α-helix, the first beta turn initiating a given stable β-hairpin cannot be positioned in any place of the chain, but *only* in its middle (Fig. 9.4). Consequently, initiation of the β-hairpin has to take $\sim n_0 \approx 30$ times

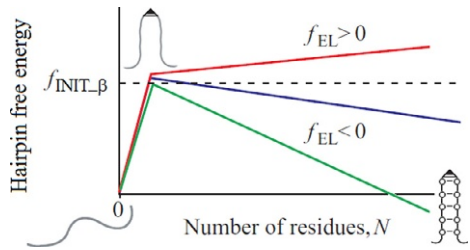


FIG. 9.4 Typical dependence of the β-hairpin free energy (ΔF_β) on the number (N) of amino acid residues forming a hairpin. A hairpin is stable only when the free energy of its elongation (f_{EL}) is negative. Note that the β-hairpin-initiating turn can be only in (approximately) the middle of the chain.

more time than initiation of the α-helix (while the β-hairpin elongation, which resembles the α-helix elongation, must take much less time) (Muñoz et al., 1998; Galzitskaya et al., 2002). This estimate shows a good fit to experiment.

Now, let's pass to the formation of β-sheets.

Theory shows that the "anomalous" (as compared with the coil-helix or coil-hairpin transition) slow kinetics of formation of the β-structure should be connected with its two-dimensionality (in contrast to a 1D helix or coil or hairpin) (Fig. 9.5), which results in a first-order phase transition.

Let us consider this process in more detail (see Finkel'shtein, 1978; Finkelstein, 1991; Galzitskaya et al., 2002).

Edge residues have fewer contacts than internal residues of the sheet. In other words, the edge of a β-sheet (like the boundary of any other phase: a drop of water, a piece of ice or an α-helix) has a higher free energy. However, like a drop of water and in contrast to the α-helix, the β-sheet is not 1D, ie, its boundary (and hence, the boundary free energy) grows with increasing number of residues involved in the sheet. Therefore, the transition from a random coil to the β-structure becomes a first-order phase transition like the formation of a water drop or a piece of ice.

Let us show that this provokes the occurrence of a high free-energy barrier (especially in the formation of an only marginally stable β-structure) capable of slowing down the folding initiation a zillion-fold.

The edge of a β-sheet consists of (a) edge β-strands, and (b) bends or loops connecting the β-strands (Fig. 9.5A). Let the coil free energy be zero (ie, the reference point); f_β, the free energy of a residue in the center of the β-sheet; $f_\beta + \triangle f_\beta$, the free energy of an edge β-strand residue (ie, $\triangle f_\beta$ is the edge effect);

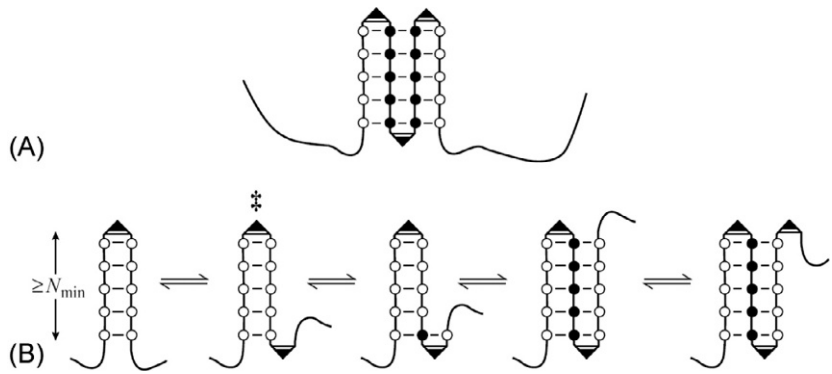


FIG. 9.5 (A) Schematic of a β-sheet. The amino acid residues of internal β-strands are indicated by *closed circles*, and of edge β-strands by *open circles*; the bends (or loops) connecting the β-strands are indicated by angles. (B) Illustration of the β-sheet growth scenario given in the text for the case when separate β-hairpins are unstable. The most unstable structure on the pathway is marked with ‡. H-bonds are shown schematically.

and U, the free energy of a bend. Since the β-sheet forms, it is stable (ie, $f_\beta < 0$), and the edge effects prevent it from falling into pieces (ie, $\triangle f_\beta > 0$ and $U > 0$).

In the kinetics of β-sheet formation we must distinguish between the following two cases:

1. $f_\beta + \triangle f_\beta < 0$, ie, in itself a long β-hairpin is more stable than a coil. Then only the turn at its top needs to overcome the activation barrier (which is almost identical to the barrier to be overcome in forming an α-helix), and subsequent growth of the β-structure is rapid, like the elongation of an α-helix or a separate β-hairpin (see the lines with $f_{EL} < 0$ in Figs. 9.3 and 9.4). This case was considered previously.
2. $f_\beta + \triangle f_\beta > 0$, ie, in itself the β-hairpin is unstable, and it is only the association of the initiating hairpin with other β-strands into a β-sheet that stabilizes the β-structure. Then the activation barrier is represented by the formation of a "nucleus," that is, such a β-sheet or β-hairpin that provides further growth of the sheet accompanied by an overall decrease of the free energy.

We will now consider just this case.

The formation and subsequent growth of the nucleus of a new phase are the most typical feature of first-order phase transitions (Landau and Lifshitz, 1980), β-structure formation among them. However, as we will see, overcoming the nucleus-provoked activation barrier may be an extremely slow process.

Let us consider the following simplest scenario of formation of a stable β-sheet when separate β-hairpins are unstable (Fig. 9.5B): (i) formation of the initiating β-hairpin by a turn and two β-strands N residues long each; (ii) formation of the next turn at its end; (iii) association of another N-residue β-strand; (iv) formation of the next turn; (v) association of another β-strand, and so on.

The formation (in a coil) of a β-hairpin consisting of a turn and two N-residue β-strands contributes as much as $U + 2N(f_\beta + \triangle f_\beta) > 0$ (because $U > 0$ and $f_\beta + \triangle f_\beta > 0$) to the free energy of the chain; formation of the next turn makes it still higher by a value of U. Association of the N-residue edge of this hairpin with a new N-residue β-strand decreases the free energy by Nf_β (since the number of edge residues remains the same, while the number of internal residues increases by N, see Fig. 9.5); formation of the next β-turn increases the free energy again by U; association of another N-residue β-strand decreases it by Nf_β, and so on.

The cycle of "association of another β-strand and formation of the next β-turn" changes the net free energy by $N\triangle f_\beta + U$. And since this cycle must result in a decrease of the free energy (as a prerequisite of rapid growth), each associating strand should contain not less than

$$N_{min} = \frac{U}{(-f_\beta)} \tag{9.10}$$

residues. It is noteworthy that this value grows to infinity when $f_\beta \to 0$, ie, when β-sheet approaches the margin of stability.

The "transition" state, ie, the most unstable state in formation of the β-structure is, according to our scenario, the β-hairpin with a subsequent turn. Since we consider the case where the hairpin stability decreases with increasing length of the hairpin, and since the β-strand of the initiating hairpin must contain at least $N_{\min}$ residues, the minimum free energy of the initiating hairpin and the next turn is

$$F^{\ddagger} = U + 2N_{\min}\left(f_{\beta} + \Delta f_{\beta}\right) + U = \frac{2U\,\Delta f_{\beta}}{\left(-f_{\beta}\right)} \tag{9.11}$$

This is the free energy of the transition state in β-sheet folding *according to our scenario* (Finkel'shtein, 1978; Finkelstein, 1991). It can be very high when f_{β} is close to zero.

Now we have to show that, irrespective of the scenario, there cannot exist transition states of a considerably higher stability.

To do so, let us estimate the changes in the minimum free energy of a growing β-sheet. The general course of the free energy changes is presented in Fig. 9.6. When the sheet is small, edge effects predominate, and the free energy of a growing β-sheet increases. In a large sheet, internal residues predominate, and the free energy of a growing β-sheet decreases.

Now we pass to calculations. First, let us estimate the minimum free energy of an M-residue β-sheet. The free energy of a sheet comprising m β-strands (which are of the same length in order to minimize the edge free energy) and $m-1$ turns is

$$F(M,m) = Mf_{\beta} + 2(M/m)\Delta f_{\beta} + (m-1)U \tag{9.12}$$

(Here I allowed myself to neglect a relatively small number of residues forming turns.) Varying the number of β-strands m (at a given M) gives the minimum of this free energy from the condition

$$\frac{\mathrm{d}F}{\mathrm{d}m} = -2\left(M/m^2\right)\Delta f_{\beta} + U \tag{9.13}$$

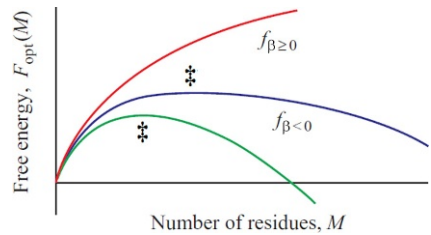


FIG. 9.6 β-Sheet minimum free energy F_{opt} as a function of M, the number of residues in the sheet. The curves refer to various free energies (f_{β}) of internal residues. The maximum of F_{opt} is marked with ‡. In a growing β-sheet (unlike an α-helix or a β-hairpin, see Figs. 9.3 and 9.4) this maximum does not correspond to the very beginning of the process.

This yields the optimal number of β-strands in this sheet, $m_{opt} = M^{1/2}(2\triangle f_\beta/U)^{1/2}$, and its free energy, $F_{opt}(M) \equiv F(M, m_{opt}) = Mf_\beta - U + 2M^{1/2}(2\triangle f_\beta U)^{1/2}$. Varying the magnitude $F_{opt}(M)$ over M gives its maximum (see Fig. 9.6) from the condition

$$\frac{dF_{opt}}{dM} = f_\beta + M^{-1/2}\left(2\Delta f_\beta U\right)^{1/2} = 0 \tag{9.14}$$

Then the sheet size corresponding to this maximum can be determined as $M^* = 2(\triangle f_\beta U)/(-f_\beta)^2$, and its free energy as

$$F^* = F_{opt}(M^*) = \frac{2U\Delta f_\beta}{\left(-f_\beta\right)} - U \tag{9.15}$$

The quantities F^* and $F^\ddagger$ (see Eq. (9.11)) coincide as to their principal term, $\frac{2U\triangle f_\beta}{(-f_\beta)}$ (Finkelstein, 1991). It is because of this term that the free energy of the transition state is *always* high when there is a low free energy of stabilization $(-f_\beta)$ of the β-structure, ie, when the β-structure is only marginally stable. (The main thing is that F^* and $F^\ddagger$ equally tend to infinity when the β-structure approaches thermodynamic equilibrium with the coil, ie, when $(-f_\beta) \to 0$).

Initiation of the β-sheet in the *given* place of the chain must take the time $t_{INIT\_\beta} \sim \tau_\beta \exp(+F^\ddagger/kT)$.

The time of β-folding initiation *somewhere* in a chain of M residues is $\sim t_{INIT\_\beta}/M$ (Finkel'shtein, 1978; Finkelstein, 1991). The expansion of the β-structure all over the chain from *one* initiation center takes $\sim M\tau_\beta$. When the chains are not extremely long, the initiation is the rate-limiting step.

For extremely long chains (and/or for intermolecular β-structures), the time of β-structure formation is about $\tau_\beta \exp(+F^\ddagger/2kT)$. (This problem can be considered in the same way as helix formation in long chains that we discussed earlier. The main idea is that in this case the number of residues (M_{eff}) included in each independently folding sheet is such that its initiation and expansion times are equal, ie, $\tau_\beta \exp(+F^\ddagger/kT)/M_{eff} = M_{eff}\,\tau_\beta$.)

Thus, in both cases, the time of β-sheet formation depends on the β-structure stability per residue, $-f_\beta$, as

$$t_\beta \sim \exp\left[\frac{A}{-f_\beta}\right] \tag{9.16}$$

when the β-structure stability is low, that is, $(-f_\beta) < \triangle f_\beta$. No matter what the numerical value of the constant A may be, it clearly follows from Eq. (9.16) that the time of β-structure formation is enormous at low $-f_\beta$ (in the limit, ie, when β-sheet approaches the margin of stability and $f_\beta \to 0$, it is *infinity*; see Fig. 9.7A).

Experiments (Snell and Fasman, 1973; Cosani et al., 1974; Walter and Fasman, 1977; Fukada et al., 1989) show a drastic dependence of the β-structure formation rate on its stability, similar to that shown in Fig. 9.7B.

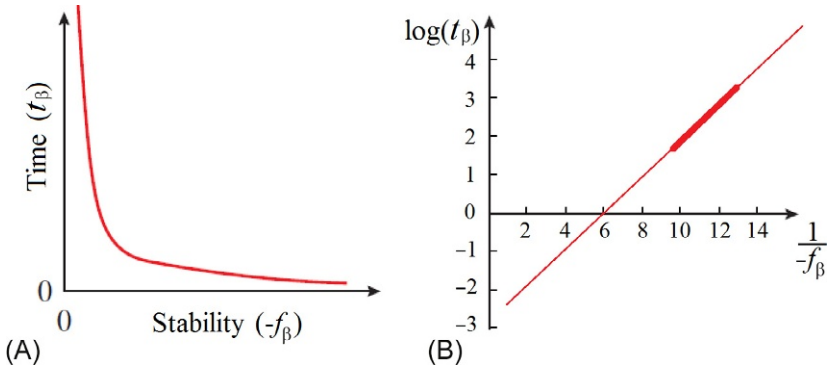


FIG. 9.7 (A) The dependence of the time needed to form a β-sheet on the stability of the β-structure (general view). (B) Time for β-structure formation (s) versus the reciprocal stability (kcal mol^{-1}). The thick line corresponds to experimental data; the thin line is its extrapolation. A high $\frac{1}{-f_\beta}$ value can be due not only to a low stability of the β-structure as related to the competing coil, but also to its low stability in relation to the competing α-helical state. *(Adapted from Finkelstein, A.V., 1991. Rate of beta-structure formation in polypeptides. Proteins 9, 23–27.)*

This explains both for the experimentally observed extremely low rate of β-structure formation in non-aggregating polypeptides (where the stability of the β-structure is always low) and for a drastic increase of this rate with increasing stability of the β-structure (Fig. 9.7B).

Thus, a β-structure of low stability must form very slowly not because of slow elongation but because of slow initiation, although β-sheets and β-hairpins of high stability (which are observed in proteins) must form almost as rapidly as the α-helix.

> *Inner voice*: Now you told us about the coil → β-structure transitions. Before, you told us about the coil → α-helix transitions. What about the α-helix → β-structure transition?
> *Lecturer*: This is a good question which needs an extended answer. I meant to say that, in all the above considerations, kinetics of the β-sheet formation was dependent on stability of the β-structure *counted off* the stability of the preceding state of the polypeptide. This "preceding state" may be the coil state in some experiments and the α-helical state in others.

When α-helices are more stable than the coil, a peculiar kinetics is observed for β-structure formation from the initial coil. It is a very fast formation of α-helices followed by a much slower α → β transition (Cosani et al., 1974). It may seem that the α-helical intermediates facilitate the β-structure formation in this case. However, this naive idea is wrong, because experiments show that a decrease in the stability of α-helices relative to the coil accelerates the β-structure formation, and a complete destabilization of α-helices and disappearance of the α-helical intermediate makes the β-structure formation even faster. This shows

that the α-helices (which, as mentioned, always fold rapidly) are *not* the on-pathway but rather *off*-pathway intermediates, and that they *do not facilitate* but rather *hinder* the β-structure formation (Finkel'shtein, 1978; Finkelstein, 1991). In other words, as the ancient Romans used to say, *post hoc non est propter hoc* (ie, "after this does not mean because of this").

A very slow initiation is a common feature of first-order phase transitions when the emerging phase is only marginally stable. Recall the overcooled liquid or the overcooled vapor … All these effects are connected with a large area and, hence, the high free energy of interface between the phases. And the β-structure is formed through a first-order phase transition with all its consequences …

In contrast, the α-helix *avoids* the first-order phase transition (remember, the helix boundary, unlike that of the β-structure (or of a piece of ice) *does not* increase with its increasing size), and therefore the barrier to be overcome in helix folding is always of a finite (and small) value; hence, the initiation here can take a fraction of a millisecond.

REFERENCES

Barskaya, T.V., Ptitsyn, O.B., 1971. Thermodynamic parameters of helix-coil transition in polypeptide chains. II. Poly-L-lysine. Biopolymers 10, 2181–2197.

Bychkova, V.E., Ptitsyn, O.B., Barskaya, T.V., 1971. Thermodynamic parameters of helix-coil transition in polypeptide chains. I. Poly-(L-glutamic acid). Biopolymers 10, 2161–2179.

Cosani, A., Terbojevich, M., Romanin-Jacur, L., Peggion, E., 1974. Potentiometric and circular dichroism studies of the coil-β transitions of poly-L-lysine. In: Proc. Symp. "Peptides, Polypeptides and Proteins", Israel, pp. 166–176.

Cummings, A.L., Eyring, E.M., 1975. Helix-coil transition kinetics in aqueous poly(α, L-glutamic acid). Biopolymers 14, 2107–2114.

Fersht, A., 1999. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman & Co., New York (Chapters 17 and 18).

Finkel'shtein, A.V., 1978. Kinetics of antiparallel β-structure formation. Bioorganicheskaya Khimiya (in Russian) 4, 340–344.

Finkelstein, A.V., 1991. Rate of beta-structure formation in polypeptides. Proteins 9, 23–27.

Finkelstein, A.V., 1995. Predicted beta-structure stability parameters under experimental test. Protein Eng. 8, 207–209.

Finkelstein, A.V., Ptitsyn, O.B., 1976. Theory of protein molecule self-organization. IV. Helical and irregular local structures of unfolded protein chains. J. Mol. Biol. 103, 15–24.

Finkelstein, A.V., Ptitsyn, O.B., Kozitsyn, S.A., 1976. Theory of protein molecule self-organization. II. A comparison of calculated thermodynamic parameters of local secondary structures with experiments. Biopolymers 16, 497–524.

Flory, P.J., 1969. Statistical Mechanics of Chain Molecules. Interscience Publishers, New York (Chapters 1–3).

Fukada, K., Maeda, H., Ikeda, S., 1989. Kinetics of pH-induced random coil-β-structure conversion of poly[S-(carboxymethyl)-cysteine]. Macromolecules 22, 640–645.

Galzitskaya, O.V., Higo, J., Finkelstein, A.V., 2002. α-Helix and β-hairpin folding from experiment, analytical theory and molecular dynamics simulations. Curr. Protein Pept. Sci. 3, 191–200.

Landau, L.D., Lifshitz, E.M., 1980. Statistical Physics (Volume 5 of A Course of Theoretical Physics), third ed. Elsevier, Amsterdam. §§ 150, 152.

Mandel, R., Fasman, G.D., 1975. The random coil-β transitions of L-lysine and L-valine: potentiometric titration and circular dichroism studies. Biopolymers 14, 1633–1650.

Muñoz, V., Serrano, L., 1994. Elucidating the folding problem of helical peptides using empirical parameters. Nat. Struct. Biol. 1, 399–409.

Muñoz, V., Thompson, P.A., Hofrichter, J., Eaton, W.A., 1997. Folding dynamics and mechanism of beta-hairpin formation. Nature 390, 196–199.

Muñoz, V., Henry, E.R., Hofrichter, J., Eaton, W.A., 1998. A statistical mechanical model for beta-hairpin kinetics. Proc. Natl. Acad. Sci. USA 95, 5872–5879.

Padmanabhan, S., Marqusee, S., Rigeway, T., Lane, T.M., Baldwin, R.L., 1990. Relative helix-forming tendencies of nonpolar amino acids. Nature 344, 268–270.

Pederson, D., Gabriel, D., Hermans Jr., J., 1971. Potentiometric titration of poly-L-lysine. The coil-to β transition. Biopolymers 10, 2133–2145.

Platzer, K.E.D., Ananthanarayanan, V.S., Andreatta, R.H., Scheraga, H.A., 1972. Helix-coil stability constants for the naturally occurring amino acids in water. IV. Alanine parameters from random poly[(hydroxypropyl)glutamine-co-L-alanine). Macromolecules 5, 177–187.

Ptitsyn, O.B., Finkel'shtein, A.V., 1970. Relation of the secondary structure of globular proteins to their primary structure. Biofizika (in Russian) 15, 757–768.

Ptitsyn, O.B., Finkelstein, A.V., 1979a. Mechanism of protein folding. Int. J. Quant. Chem. 16, 407–416.

Ptitsyn, O.B., Finkelstein, A.V., 1979b. A problem of protein structure prediction. In: Volkenstein, M.V. (Ed.), 'Itogi Nauki I Techniki', Ser. 'Mol. Biologiya' ('Results of Science and Technology', Ser. 'Mol. Biology'; in Russian), vol. 15. VINITI, Moscow, pp. 6–41.

Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York (Chapters 1 and 2, Appendix).

Snell, C.R., Fasman, G.D., 1973. Kinetics and thermodynamics of the helix leads to transconformation of poly(L-lysine) and L-leucine copolymers. A compensation phenomenon. Biochemistry 12, 1017–1025.

Thompson, P.A., Eaton, W.A., Hofrichter, J., 1997. Laser temperature jump study of the helix-coil kinetics of an alanine peptide interpreted with a "kinetic zipper" model. Biochemistry 36, 9200–9210.

von Dreele, P.H., Lotan, N., Ananthanarayanan, V.S., et al., 1971. Helix-coil stability constants for the naturally occurring amino acids in water. II. Characterization of the host polymers and application of the host-guest technique to the random poly(hydroxipropylglutamine)-co-(hydroxibutylglutamine). Macromolecules 4, 408–417.

Walter, B., Fasman, G.D., 1977. The random coil-β transitions of copolymers of L-lysine and L-isoleucine: potentiometric titration and circular dichroism studies. Biopolymers 16, 17–32.

Williams, S., Causgrove, T.P., Gilmanshin, R., Fang, K.S., Callender, R.H., Woodruff, W.H., Dyer, R.B., 1996. Fast events in protein folding: helix melting and formation in a small peptide. Biochemistry 35, 691–697.

Wooley, S.Y., Holzwarth, G., 1970. Intramolecular β-pleated-sheet formation by poly-L-lysine in solution. Biochemistry 9, 3604–3608.

Zana, R., 1975. On the rate determining step for helix propagation in the helix-coil transition of polypeptides in solution. Biopolymers 14, 2425–2428.

Zimm, B.H., Bragg, J.K., 1959. Theory of phase transition between helix and random coil in polypeptide chains. J. Chem. Phys. 31, 526–535.

This page intentionally left blank

Lecture 10

Now, let us discuss the properties of the side chains of amino acid residues. In particular, I would like to consider the question of what structures stabilize individual residues.

The list of 20 "standard" DNA-encoded amino acid residues (Cantor and Schimmel, 1980; Lehninger et al., 1993) is given in Table 10.1, and structures of their side chains are presented in Fig. 10.1, together with structures of two "nonstandard," though equally DNA-encoded, amino acid side chains.

Apart from 20 standard amino acid residues shown in the right part of Fig. 10.1, there are some rare nonstandard ones. Most of them are produced by modifications of standard amino acids, but two, as has become known more or less recently, namely, "selenocysteine" and "pyrrolysine" (see the right part of Fig. 10.1) are coded by some RNA codons in some organisms. These are stop-codons UGA for selenocysteine and UAG for pyrrolysine positioned in special RNA contexts (Creighton, 1993; Longtin, 2004; Srinivasan et al., 2002).

Also, *N*-formylmethionine (fMet) amino acid (which is a derivative of Met with a formyl group –COH added to its amino group) is a gene-encoded residue that is used for initiation of protein synthesis in some cases. But fMet has the same side chain as Met.

Let us consider the structural tendencies of amino acid residues: these have become known after long-term statistical investigations of protein structures. Such investigations answer the question as to what is most likely to happen and what is not.

Table 10.2 may be helpful in putting these answers in order. Along with the abundance in different parts of proteins, I have tabulated such residue properties as the presence of an NH group in the main chain (it is absent only from the imino acid proline), the presence of the C^{β}-atom in the side group (it is absent only from glycine), the number of nonhydrogen γ-atoms in the side chain, and the presence and type of polar groups in the side chain (dipoles or charges with a sign; the charged state corresponding to a "normal" pH of 7.0 is shown in bold).

Let us try and understand the major features of Table 10.2, based on what we have already learned. In doing so, we will follow a logical criterion "what's good for General Motors is good for America": since the protein as a whole is stable, the majority of its components must be stable, that is, stable components must be most often observed in its structure, while nonstable ones must be rare.

Why does proline dislike the secondary structures? Because it lacks the main-chain NH group, that is, its ability of H-bonding is halved, and H-bonds are of primary importance for the secondary structure. Why does it nevertheless like the N-terminus of the helix? Because here, at the N-terminus, NH groups

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00010-7

TABLE 10.1 The Principal Properties of 20 Standard Natural Amino Acid Residues

Amino Acid Residue	Code		Occurrence in *E. coli* Proteins (%)	MW at pH 7 (Da)	$G_{water \rightarrow alcohol}$ of side chain at 25°C (kcal mol^{-1})
	3-Letter	*1-Letter*			
Glycine	Gly	G	8	57	0
Alanine	Ala	A	13	71	−0.4
Proline	Pro	P	5	97	−1.0
Glutamic acid	Glu	E	≈6	128	+0.9
Glutamine	Gln	Q	≈5	128	+0.3
Aspartic acid	Asp	D	≈5	114	+1.1
Asparagine	Asn	N	≈5	114	+0.8
Serine	Ser	S	6	87	+0.1
Histidine	His	H	1	137	−0.2
Lysine	Lys	K	7	129	+1.5
Arginine	Arg	R	5	157	+1.5
Threonine	Thr	T	5	101	−0.3
Valine	Val	V	6	99	−2.4
Isoleucine	Ile	I	4	113	−1.6
Leucine	Leu	L	8	113	−2.3
Methionine	Met	M	4	131	−1.6
Phenylalanine	Phe	F	3	147	−2.4
Tyrosine	Tyr	Y	2	163	−1.3
Cysteine	Cys	C	2	103	−2.1
Tryptophan	Trp	W	1	186	−3.0

All data are from Schulz and Schirmer (1979/2013), except for those on side chain hydrophobicity ($G_{water \rightarrow alcohol}$), which are from Fauchére and Pliska (1983). The volume (in Å^3) occupied by a residue (in protein or in water) is close to its molecular weight (in Da) multiplied by 1.3. To be more precise, it is ∼5% higher than MW × 1.3 if the residue contains many aliphatic (–CH_2–, –CH_3) groups and ∼5% lower than MW × 1.3 if the residue contains many polar (O, N) atoms.

protrude from the helix, that is, they do not participate in hydrogen bonds, and proline loses nothing here (the same refers to definite positions at the β-sheet edges). In addition, angle φ is close to −60 degree in the helix, while in proline, its ring fixes the angle φ at about −60 degree, that is, proline is about ready to adopt the helical conformation (Fig. 10.2A).

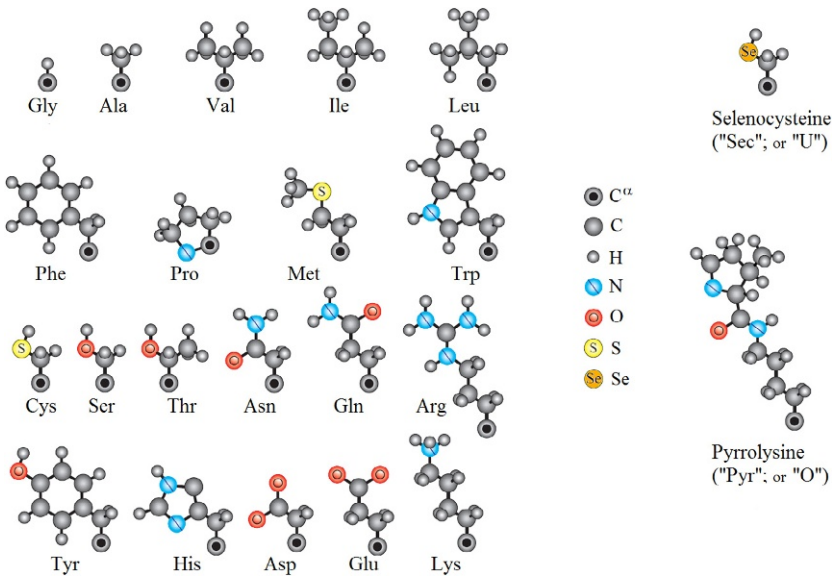


FIG. 10.1 The side chains of 20 standard (on the left) and two rare "nonstandard" (on the right) amino acid residues. Atoms involved in the amino acids are shown in the center.

Why does glycine dislike the secondary structure and prefer irregular segments (coil)? Because its allowed $\varphi\psi$ region in the Ramachandran map is extremely broad (Fig. 10.2B), and it can easily adopt a variety of conformations other than secondary structure.

In contrast, alanine with its more narrow (but including both α and β conformations) region of allowed conformations (Fig. 10.2B) prefers the α-helix (and partially the β-structure) rather than irregular conformations.

Other hydrophobic residues (ie, residues without charges and dipoles in their side chains) prefer, as a rule, the β-structure. Why? Because there is more room for their large γ-atoms (Fig. 10.2C). This is of particular importance for "branched" side chains with two large γ-atoms (Leach et al., 1966), and indeed, these are strongly attached to the β-structure.

As to amino acids with polar groups in their side chains, they prefer irregular (coil) surface regions where these polar groups can easily participate in H-bonds with both the irregular polypeptide chain and water. This tendency is most clearly displayed by most polar residues, which are charged at a "normal" pH of 7.0, as well as by the shortest polar side chains whose polar groups are closest to the main chain. By the way, this possibility of additional H-bonding explains the tendency of short polar side chains to be located at both ends of the helix.

Tryptophan and tyrosine are a kind of exception from amino acids having dipoles in their side chains: they have a small dipole and a large hydrophobic part; another exception is cysteine whose SH groups make extremely weak H-bonds. Their behavior is rather similar to that of hydrophobic residues.

TABLE 10.2 The Principal Structural Properties of Amino Acid Residues

Residue	Main Chain[a] NH	Side Chain[a] *Number of C^{β}*	Side Chain[a] *Number of γ*	Dipole/Charge[b]	pK_a[b]	Structural Occurrence Tendency[c] *Before* α_N	*In helix* α_N	*In helix* α	*In helix* α_C	*After* α_C	*In* β	*Loops*	*Core*
Gly	+	0	0				−				−	+	
Ala	+	1	0				+					−	
Pro	No	1	1				+	−	−	−	−	+	
Glu	+	1	1	COOH → $\mathbf{CO_2^-}$	4.3	+	+		−	−	−		−
Asp	+	1	1	COOH → $\mathbf{CO_2^-}$	3.9	+	+	−	−	−	−	+	−
Gln	+	1	1	$OCNH_2$									−
Asn	+	1	1	$OCNH_2$		+		−		+	−	+	−
Ser	+	1	1	OH		+						+	−
His	+	1	1	NH; &**N** → NH^+	6.5		−		+	+			
Lys	+	1	1	$NH_2 \rightarrow \mathbf{NH_3^+}$	10.5	−	−		+	+	−		−
Arg	+	1	1	$\mathbf{HNC(NH_2)_2^+}$	12.5	−	−		+	+	−	+	−
Thr	+	1	2	OH		+					+		
Ile	+	1	2								+	−	+
Val	+	1	2								+	−	+
Leu	+	1	1					+			+	−	+

Met	+	1	1					+			+	–	+
Phe	+	1	1								+	–	+
Tyr	+	1	1	OH → **O**$^-$	10.1			–			+		+
Cys	+	1	1	**SH** → S$^-$	9.2			–			+		+
Trp	+	1	1	NH							+		+

[a]*See text for a definition of these factors.*
[b]*Bold type in the "dipole/charge" column shows the state of the ionizable group at the "neutral" pH 7; pK is that pH value where the group can be in the charged and uncharged states with equal probability (see Fig. 10.5).*
[c]*These columns refer, respectively, to the tendency to be: immediately before the helix N-terminus; in the α-helix (the N-terminal turn, the body and the C-terminal turn); immediately after the C-terminus; in the β-structure; in irregular structures (loops), including β-turns of the chain; and in the hydrophobic core of the globule, rather than on its surface. The "tendency" is measured as a concentration of a residue in a given structure relatively to the average concentration of this residue in studied proteins. The difference in occurrence between "+" and "–" is approximately a factor of two. A particularly strong tendency is shown by a larger, bold symbol.*
Data from Ptitsyn, O.B., Finkel'shtein, A.V., 1970. Relation of the secondary structure of globular proteins to their primary structure. Biofizika 15, 757–768 (in Russian); Ptitsyn, O.B., Finkelstein, A.V., 1979. A problem of protein structure prediction. In: Volkenstein, M.V. (Ed.), Itogi Nauki I Techniki (Results of Science and Technology). Mol. Biologiya (Mol. Biology), vol. 15, VINITI, Moscow, pp. 6–41 (in Russian); Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York, NY/Heidelberg/Berlin; Miller, S., Janin, J., Lesk, A.M., Chothia, C., 1987. Interior and surface of monomeric proteins. J. Mol. Biol. 196, 641–656; Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman & Co., New York (Chapter 5); Stepanov, V.M., 1996. Molecular Biology: Protein Structure and Function. Vysshaya Shkola, Moscow (Chapter 5, in Russian).

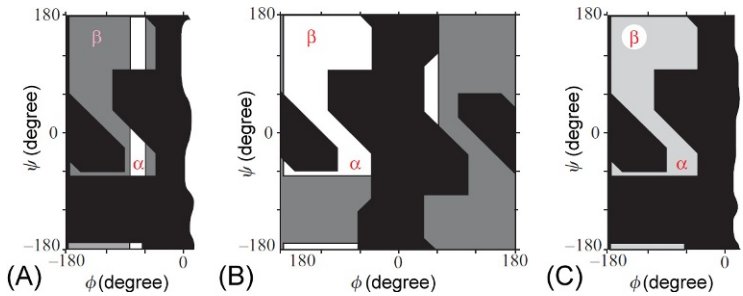


FIG. 10.2 Disallowed and allowed conformations of various amino acid residues as a background for α and β-conformations. (A) Allowed proline conformations (□) against allowed alanine conformations ▩; ■, conformations disallowed for both of them. (B) Allowed alanine conformations (□) against conformations allowed for glycine only ▩; ■, conformations disallowed for all residues. (C) The map of disallowed ■ and allowed (□, ▩) conformations of larger residues. □, the region where all side-chain χ^1 angles are allowed; ▩, the region where some χ^1 angles are disallowed. (See Lectures 3 and 7 for discussion of these Ramachandran-like plots.)

As it has been observed (Ptitsyn and Finkel'shtein, 1970; Ptitsyn and Finkelstein, 1979) that the negatively charged side chains have a preference for the N-terminus of the helix (or rather, to the N-terminal turn plus one or two preceding residues). At the same time, they avoid the C-terminal turn (plus one or two subsequent residues). The preference of the positively charged groups is just the opposite. What is the reason for that? It is NH groups protruding from the N-terminus and creating a considerable positive charge that attracts "minuses" and repels "pluses" of the side chains (Fig. 10.3). In contrast, the C-terminus is charged negatively, and therefore it is attractive for "pluses" of the side chains, while their "minuses" avoid it.

As to the residue location in the globule, the general tendency is that polar (hydro*philic*) side chains are on the protein surface, in contact with polar water molecules (like dissolves in like). Separation from water molecules negatively affects polar groups since then they lose their H-bonds. This negative effect is especially profound when it concerns the charged groups because their transition from the high permittivity medium (water) to that of low permittivity (protein core) is accompanied by a drastic increase in the free energy. Indeed, ionized groups are almost completely absent from the protein interior (nearly all exceptions are connected with either coordinate bonding of metal ions or with active sites which are, in fact, the protein's focus…).

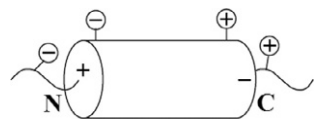


FIG. 10.3 Favorable positions of charged side chains near the N- and C-ends of the helix (Ptitsyn and Finkel'shtein, 1970; Ptitsyn and Finkelstein, 1979).

In contrast, the majority of hydrophobic side chains are in the protein interior to form the hydrophobic core (again, "like dissolves in like"). As we have learned, the hydrophobicity of a group grows with its nonpolar surface that is to be screened from water. For purely nonpolar groups, the hydrophobic effect is directly proportional to their total surface, while in the case of polar admixtures it is proportional to their surface less the surface of these admixtures.

Adhesion of hydrophobic groups is the main, although not the only, driving force of protein globule formation: it is also assisted by H-bonding in the secondary structure (as discussed earlier) and by tight quasicrystal packing within the interior of the protein molecule (to be discussed later).

To form the protein hydrophobic core, the chain must enter it with hydrogen bonds already formed (or formed in the process) because the rupture of H-bonds between polar peptide groups and water molecules occurring in any other way is expensive. That is why the chain involved in hydrophobic core formation has already formed its secondary structure (or forms it in the process), and thereby saturates the hydrogen bonds of the main-chain peptide groups.

The core must mostly comprise hydrophobic side chains from secondary structures, while polar side chains of the same secondary structures must remain outside; therefore, both α-helices and β-strands located at the surface have hydrophobic and hydrophilic surfaces created by alternating appropriate groups of the protein chain in a certain order (Fig. 10.4).

All the regularities just discussed are used to build artificial (*de novo*) proteins and to predict the secondary structure of proteins from their amino acid sequences, and also to predict internal and surface portions of protein sequence segments either deeply immersed in the protein or positioned on its surface. We will discuss this later.

In conclusion, a few more words about ionized side groups. Increasing pH (ie, decreasing H^+ concentration) always shifts the charged state of a group "in the negative direction," that is, a neutral group acquires a negative charge, and a positive group becomes discharged (see Fig. 10.5).

Different groups change from the uncharged to the charged state or vice versa at different pH, but the transition width always remains the same—about 2 units of pH (within this range the charged/uncharged state ratio changes from 10:1 to 1:10).

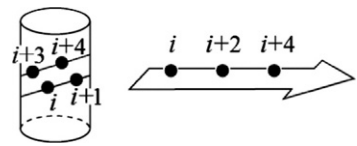


FIG. 10.4 The positions at which nonpolar side groups can form continuous hydrophobic surfaces on α-helices and β-strands. The numbers show residue positions in the chain. Similar combinations of polar groups result in the formation of hydrophilic areas on opposite surfaces of α-helices and β-strands (Lim, 1974a,b).

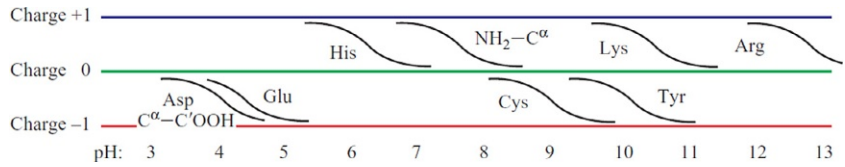


FIG. 10.5 The polarization of ionized side groups, as well as the N- and C-termini of the polypeptide chain (NH_2–C^{α} and C^{α}–C′OOH, respectively) in water at various pH values. The pH value corresponding to the "half-charged" state of a group is the p*K* value of its ionization (cf. Table 10.2). A quickly installed equilibrium of two (with an H and without it) forms of acids and bases shows that transition of an H atom from water to acid or base is relatively easy, despite the chemical nature of the covalent bonds O–H and N–H. The ratio of probabilities of charged and uncharged states is $10^{(pK-pH)}$: 1 for a positively charged group, and $10^{-(pK-pH)}$: 1 for a negatively charged group. *(Data from Stepanov, V.M., 1996. Molecular Biology: Protein Structure and Function. Vysshaya Shkola, Moscow (Chapter 5, in Russian).)*

Special attention should be paid to groups that change their uncharged state to a charged one at a pH of about 7.0 typical of proteins in a living eukaryotic cell; these easily rechargeable groups (histidine in particular) are often used in protein active sites.

As it has been already mentioned, an ionizable group easier penetrates into a nonpolar medium (eg, protein or membrane interior) in its uncharged form. Indeed, the estimated cost of an ion penetration into such a medium is as high as several dozen kcal mol^{-1}. And what does the discharging cost? This can be easily estimated from Fig. 10.5. The probability of uncharged state is $W_0 = 1/[1+10^{(pK-pH)}]$ for a positively charged group, and $W_0 = 1/[1+10^{-(pK-pH)}]$ for a negatively charged group in water. Thus, the free energy of uncharging is $F_0 = -kT \ln W_0$. That is, for a positively charged group $F_0 \approx 0$ at pH > p*K*, and $F_0 \approx 2.3kT(pK - pH)$ at pH < p*K*; for a negatively charged group $F_0 \approx 0$ at pH < p*K*, and $F_0 \approx -2.3kT$ $(pK - pH)$ at pH > p*K*. Thus, the free energy of discharging does not exceed several kcal mol^{-1} (at a "normal" pH ≈ 7) for all the ionizable groups shown in Fig. 10.5.

REFERENCES

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman & Co., New York, NY (Part 1, Chapter 2).

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman & Co., New York, NY (Chapter 5).

Fauchére, J.L., Pliska, V., 1983. Hydrophobic parameters of amino acid side chains from the partitioning of N-acetyl-amino acid amides. Eur. J. Med. Chem. Chim. Ther. 18, 369–375.

Leach, S.J., Némethy, G., Scheraga, H.A., 1966. Computation of the sterically allowed conformations of peptides. Biopolymers 4, 369–407.

Lehninger, A.L., Nelson, D.L., Cox, M.M., 1993. Principles of Biochemistry, second ed. Worth Publishers, New York (Chapter 5).

Lim, V.I., 1974a. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J. Mol. Biol. 88, 857–872.

Lim, V.I., 1974b. Algorithm for prediction of α-helices and β-structural regions in globular proteins. J. Mol. Biol. 88, 873–894.

Longtin, R., 2004. A forgotten debate: Is selenocysteine the 21st amino acid? J. Natl. Cancer Inst. 96, 504–505.

Ptitsyn, O.B., Finkel'shtein, A.V., 1970. Relation of the secondary structure of globular proteins to their primary structure. Biofizika 15, 757–768 (in Russian).

Ptitsyn, O.B., Finkelstein, A.V., 1979. A problem of protein structure prediction. In: Volkenstein, M.V. (Ed.), Itogi Nauki I Techniki. Results of Science and Technology. Mol. Biologiya (Mol. Biology), vol. 15. VINITI, Moscow, pp. 6–41 (in Russian).

Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York, NY/Heidelberg/Berlin (Chapter 5).

Srinivasan, G., James, C.M., Krzycki, J.A., 2002. Pyrrolysine encoded by UAG in Archaea: Charging of a UAG-decoding specialized tRNA. Science 296, 1459–1462.

This page intentionally left blank

Part IV

Protein Structures

This page intentionally left blank

Lecture 11

Now that we know the features of polypeptide secondary structures and the properties of amino acid residues, we can, at last, pass to proteins.

The "living conditions," structure-stabilizing interactions and overall architecture of proteins provide the basis for classifying them as (1) fibrous proteins; (2) membrane proteins; (3) water-soluble globular proteins; and (4) natively disordered (or "intrinsically disordered") proteins.

In this lecture, we will consider fibrous proteins. Their structure is simpler than that of other well-ordered proteins (Volkenstein, 1977; Schulz and Schirmer, 1979/2013; Cantor and Schimmel, 1980; Creighton, 1993; Lehninger et al., 1993; Stryer, 1995; Stepanov, 1996; Branden and Tooze, 1999).

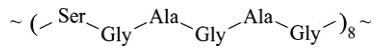
The function of fibrous proteins is mostly structural. They form microfilaments fibrils, hair, silk, and other shielding textures; they reinforce membranes and maintain the structure of cells and tissues. Fibrous proteins are often very large. Among them, there is the largest known protein, titin, of about 30,000 amino acid residues.

Fibrous proteins often form enormous aggregates; their spatial structure is mostly highly regular, usually composed of huge secondary-structure blocks, and reinforced by interactions between adjacent polypeptide chains. The primary structure of fibrous proteins is usually characterized by high regularity and periodicity, which ensures the formation of vast regular secondary structures.

We shall consider some typical representatives of fibrous proteins.

(a) *β-Structural proteins like silk fibroin*. As we know, periodicity of a β-sheet is manifested by residues pointing alternately above and below the sheet (Fig. 11.1).

In silk fibroin, the major motif of the primary structure is an octad repeat of six residue blocks, each consisting of alternating smaller (Gly) and larger residues, for example,



and this octad repeat occurs about 50 times, separated by less regular sequences (Stepanov, 1996).

Antiparallel (such as those in Fig. 11.1) β-sheets of silk fibroin are placed onto one another in the "face-to-face, back-to-back" manner (Dickerson, 1964): a double sheet of glycines (the distance between the planes is

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00011-9

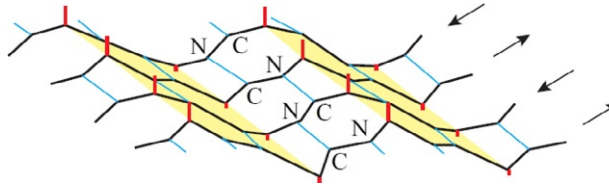
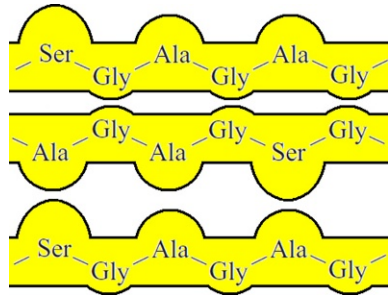


FIG. 11.1 A β-sheet with its pleated structure and periodicity emphasized. Hydrogen bonds between the linked β-strands are shown in light blue; the distance between the β-strands is 4.8 Å. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York/Heidelberg, Berlin (Chapter 5), with permission.)*

3.5 Å)—a double sheet of alanines/serines (as clearly seen by X-rays, the distance between the planes is 5.7 Å)—a double sheet of glycines—and so on:



In a silk fiber, these quasicrystals consisting of many β-sheets are immersed in a less-ordered matrix formed by irregular parts of fibroin, as well as by sericin, a special disordered matrix protein, that is S–S-bonded into a huge aggregate.

(b) α-*Structural fibrous proteins formed by long coiled-coil helices* (Fig. 11.2). In α-keratin, in all proteins of intermediate filaments (that have very different primary structures), in tropomyosin such helices cover the entire protein chain, and the major part of the myosin chain also forms a fibril of this type. These structures are also observed in some silks (not in the silkworm product considered above but in silk produced by bees and ants).

Associated helical chains form a superhelix known as a "coiled coil" (Figs. 11.2 and 11.3) (Crick, 1953; Dickerson, 1964).

FIG. 11.2 Right-handed coiled-coil α-helices. In the complex they are parallel and slightly wound around each other to form a left-handed supercoil with a repeat of 140 Å. The interhelical contacts are formed by amino acid residues at repeating chain positions **a** and **d** (see Figs. 11.3 and 11.4).

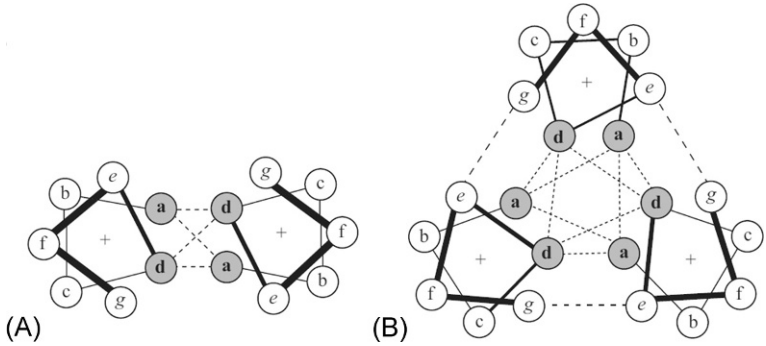


FIG. 11.3 Interactions of α-helices in double (A) and triple (B) superhelices (as viewed along the helix axis). In the double helix, only residues **a** and **d** are in immediate contact with another helix, while in the triple helix *e* and *g* residues are also involved in contacts (although to a lesser extent). *(Adapted from Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman & Co., New York (Chapter 5).)*

··· - **a** - b - c - **d** - *e* - f - *g* - **a** - b - c - **d** - *e* - f - *g* - **a** - b - c - **d** - *e* - f - *g* - ···
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

FIG. 11.4 Typical 7-residue repeats in primary structures of α-supercoil-forming chains.

The coiled coil is usually formed by parallel α-helices. In different proteins there may be two, three, or more α-helices forming the coiled coils.

As we know, a regular α-helix has 3.6 residues per turn, while the residue repeat of coiled-coil helices is 7.0, that is, 3.5 residues per turn (Figs. 11.3 and 11.4). The typical primary structure of a supercoil-forming chain has the same 7.0 residue repeat (Fig. 11.4; here, lettering in bold corresponds to hydrophobic amino acids forming the main interhelical contacts, while other letters refer to hydrophilic amino acids).

Interestingly, a slight increase in the hydrophobicity of "intermediate" residues *e* and *g* turns the double supercoiled helix (Fig. 11.3A) into a triple one (Fig. 11.3B), a greater increase turns it into a quadruple helix, and so on.

The next higher structural level is the association of supercoiled helices (shown in Fig. 11.2) into fibrils; this happens often, though not always, for example, it happens in myosin but not in tropomyosin.

It is also of interest that mechanical tension of a wet fiber formed by α-helices may result in its transformation into the β-structure, and when the tension is released and moisture decreased, the α-helical structure restores (Kreplak et al., 2004).

Let us consider in more detail how the helices associate (Crick, 1953; Chothia et al., 1977). Crick observed that there are regular knobs (side chains) on the surface of an α-helix with regular holes between them and suggested that "knob-to-hole" complementarity provides some privileged angles of contact between helices stuck together. This was a crude but correct picture of

helix-with-helix interaction. Subsequent refinement (Chothia et al., 1977) pointed out that the α-helix has several "ridges" formed by side groups coming close together (Fig. 11.5A), partly due to existence of privileged side-chain rotamers. The periodicity of some of these ridges is of the 1–4–7–⋯ type (ridges "$i, i+3$," or simply "+3"). The other ridges have periodicity of the 0–4–8–12–⋯ type (ridges "$i, i+4$," or simply "+4"). The helix-to-helix contact-zone-involved parts of ridges of the former type consist of residue pairs $\mathbf{a}_1$–$\mathbf{d}_4$, $\mathbf{a}_8$–$\mathbf{d}_{11}$,⋯ (Fig. 11.5B), while those of ridges of the latter type consist of the residue pairs $\mathbf{d}_4$–$\mathbf{a}_8$, $\mathbf{d}_{11}$–$\mathbf{a}_{15}$,⋯ (Fig. 11.5C).

The angle between the helix axis and the "$i, i+4$" ridges is about –25 degree, that between the axis and the "$i, i+3$" ridges is about +45 degree (Fig. 11.5).

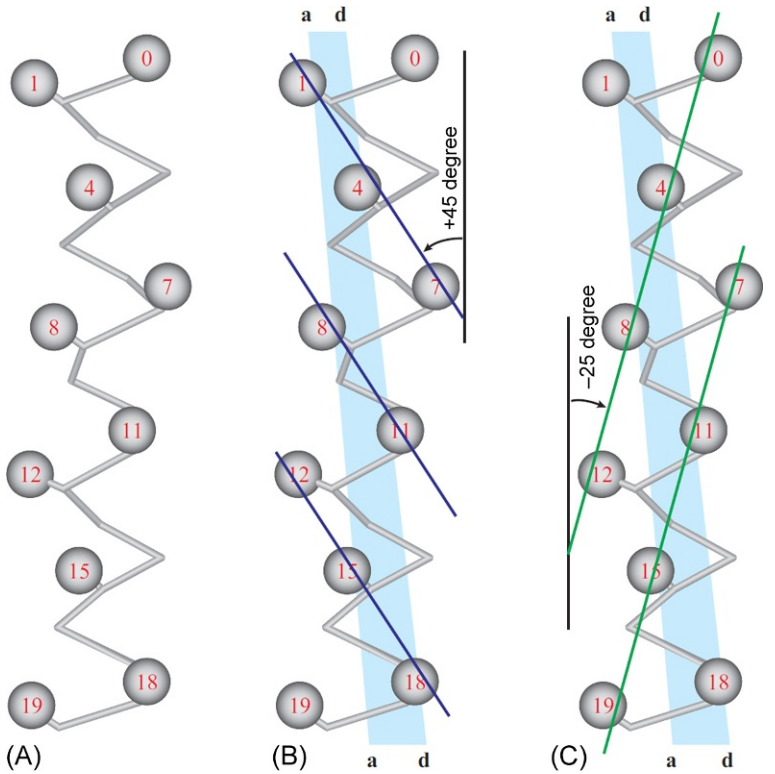


FIG. 11.5 (A) The α-helix (main chain and C^{β}-atoms). (B and C) Two types of ridges (thin lines) formed by side groups on its surface that are close in space. (B) Ridges of "$i, i+3$" type; (C) ridges of "$i, i+4$" type. The bands show the "contact surface" acting in the coiled coil; its edges are formed by the lines of residues **a** and **d**. Typical angles between the helix axis and ridges $i, i+3$ and $i, i+4$ are given (the angles in the figure look smaller than the typical angles given, since the typical ridges run, not through centers of C^{β}-atoms, but through massive side groups). *(Adapted from Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second ed. Garland Publishing Inc., New York/London (Chapter 3).)*

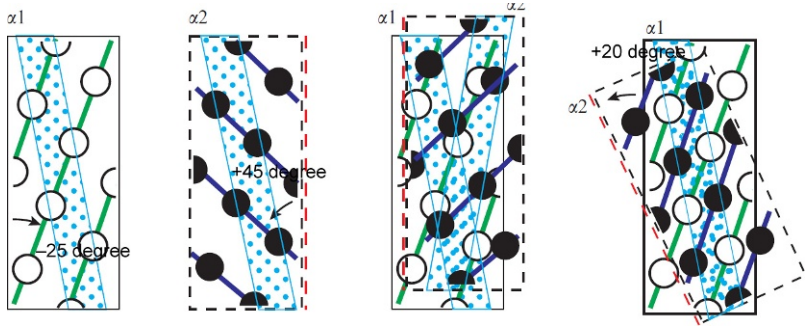


FIG. 11.6 To ensure close packing of helix side-chain ridges, a +20 degree turn of one helix about the other is required. The contact area is viewed through the α2 helix turned around its axis. Residues of the "lower" (α1) helix are light circles, of the "upper" (α2) helix, dark circles. The lines of contact for the residues **a** and **d** are shown in light blue. *(Adapted from Chothia, C., Levitt, M., Richardson, D., 1977. Structure of proteins: packing of α-helices and pleated sheets. Proc. Natl. Acad. Sci. U. S. A. 74, 4130–4134.)*

If one helical surface is turned about the vertical axis and superimposed on the other surface (Fig. 11.6) with a subsequent turn by +20 degree around the axis perpendicular to the plane of the picture, the 1–4–7 ridges of one helix will fit between the 0–4–8 ridges of the other helix (and vice versa) and ensure a close contact between the helices (Fig. 11.6, right). Then **a**-groups of one helix fit between **d**-groups of the other, and a slightly twisted contact line arises on the surface of each helix. And when these helices become intertwined (Fig. 11.2) and coiled around the common axis, the interaction area becomes straight and appears in the middle of a slightly twisted helical bundle.

This is not the only way to ensure a close contact of helices; others will be considered when discussing globular proteins. But this is the only good way for the very long helices typical of fibrous proteins. It was predicted by Crick (1953), the same year that he and Watson predicted the DNA double helix.

(c) *Collagen* ("glue forming," in Greek). This is the major structural protein amounting to a quarter of the total mass of proteins of vertebrates. It forms strong nonsoluble fibrils. A collagen molecule is formed by a special superhelix consisting of three polypeptides (Fig. 11.7) that are free of intra-chain H-bonding (Traub and Piez, 1971) and supported by some interchain hydrogen bonds only.

The conformation of all residues of each collagen chain is close to that of a polyproline (or rather, a poly(Pro)II helix), but collagen chains form additional interchain H-bonds. This helix is a left-handed helix with a 3.0 residue repeat. Accordingly, in collagen, the main motif of the primary structure is the repeated triad of residues $(\text{Gly-Pro-Pro})_n$ or rather $(\text{Gly-something-Pro})_n$. Gly is essential for hydrogen bonding in collagen since it (unlike Pro) has an NH group and no side chain; and any side chain would be unwanted in the middle of a tight collagen superhelix where the H-bonding glycine is positioned.

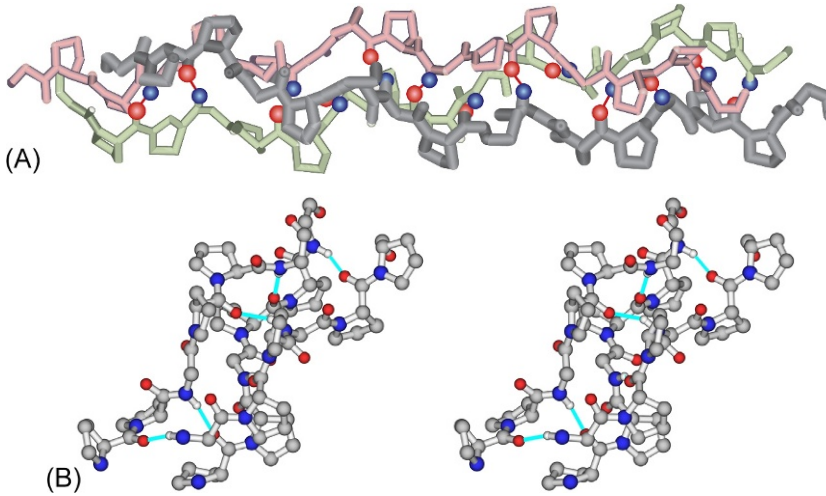


FIG. 11.7 (A) A model of the triple collagen helix with the $(Gly\text{-}Pro\text{-}Pro)_n$ repeat. Each chain is colored differently. H-atoms of NH-groups of glycines (blue) and O-atoms of the first proline in the Gly-Pro-Pro repeat (red) are shown to participate in hydrogen bonds. Gly of chain "1" binds to chain "2," while its Pro binds to chain "3," and so on. Each chain is wound around two others forming a right-handed superhelix. It is called a *super*helix because at a lower structural level, the level of conformation of separate residues, each individual collagen chain is already helical (it forms a *left-handed* "microhelix" of the poly(Pro)II type, with three residues per turn, which is easily seen from the alignment of the proline rings). A collagen molecule is about 15 Å wide and about 3000 Å long. (B) A stereo drawing of the triple collagen superhelix.

Interestingly, the collagen chain-coding exons almost always start with glycines, and the number of their codons is a multiple of three. As you may remember, eukaryotic genes contain protein-coding exons separated by introns that are cleaved from mRNA (and therefore do not encode proteins) (Creighton, 1993; Stepanov, 1996).

At the next higher structural level collagen superhelices associate into collagen fibrils.

The biosynthesis of collagen, its subsequent modification and the formation of the mature structure of a collagen fibril have been well studied (Fig. 11.8) (Schulz and Schirmer, 1979/2013; Creighton, 1993; Stepanov, 1996).

As with folding of many fibrous (and membrane) proteins, collagen folding is an externally assisted process. This distinguishes it from the most interesting *spontaneous* folding typical of water-soluble globular proteins, which is to be discussed close to the end of this course. Besides, collagen folding always involves several chains, which is typical of fibrous proteins, unlike membrane and water-soluble globular proteins.

Correct collagen folding needs to be initiated by *pro*collagen that contains, apart from collagen chains, globular heads and tails. Without these heads and tails, the collagen chains fold into "incorrect" triple helices that lack the

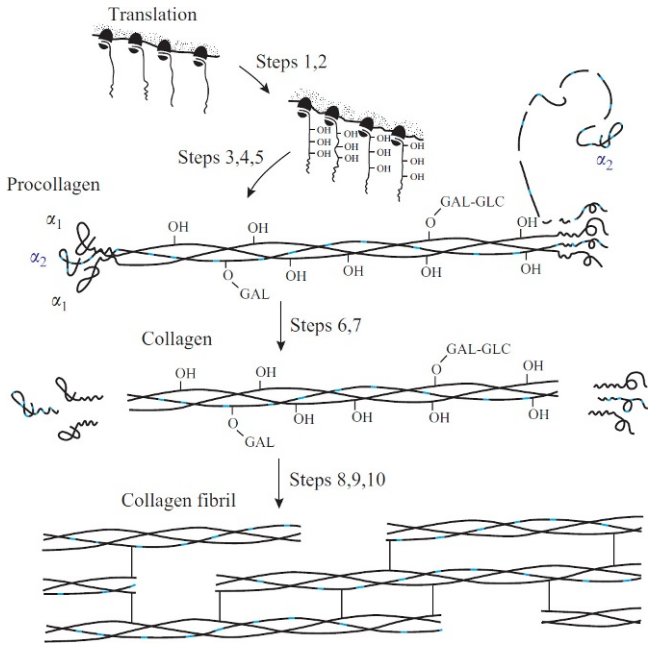


FIG. 11.8 In vivo formation of collagen. Step 1. Biosynthesis of pro-α1-chains and pro-α2-chains (≈1300 residues long each) in the ratio of 2:2. Step 2. Enzymatic cotranslational hydroxylation of some Pro and Lys residues positioned prior to Gly. Step 3. Association of sugars (GLC-GAL) with some hydroxylated residues. Step 4. Formation of the tetramer from C-terminal globules of two pro-α1 and two pro-α2 chains; subsequent degradation of one pro-α2 chain and formation of the procollagen (pro-α1-α2-α1 trimer) with S–S-bonds between globular ends. Step 5. Formation of a triple helix in the middle of procollagen. Step 6. Procollagen secretion into extracellular space. Step 7. Cleavage of globular parts. Steps 8–10. Spontaneous formation of fibrils from the triple superhelices, final modification of amino acid residues and crosslinking (caused by a special enzyme) between modified residues of collagen chains. *(Adapted from Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York/Heidelberg, Berlin (Chapter 4), with permission.)*

heterogeneity typical of native collagen (which contains one α2 and two α1 chains), its inherent register (ie, the correct shift of chains about one another), etc. Thus, a separately taken collagen triple helix is incapable of spontaneously *self*-organizing its correct spatial structure in vitro. In this respect, it is similar to silk fibroin and distinct from some of the previously described α-helical coiled coils and especially from globular proteins that we will discuss later.

As temperature is increased, the collagen helix melts (this is how gelatin is formed). The collagen melting temperature is strongly dependent on the proportion of proline and oxyproline (the higher the concentration of these rigid residues, the higher the melting temperature, naturally). This melting temperature is usually only a few degrees higher than the body temperature of the host

animal (Alexandrov, 1965). Please note this fact, as we will return to it in a future lecture.

Collagen folding is particularly interesting because a number of hereditary diseases are known to be associated with mutations in collagens (Stepanov, 1996), the best characterized of which is *osteogenesis imperfecta* or "brittle bone" disease. The most common cause of this syndrome is a single base substitution that results in the replacement of glycine by another amino acid. This breaks the $(Gly\text{-}X\text{-}Y)_n$ repeating sequence that gives rise to the characteristic triple-helical structure of collagen. It appears that at least one effect of this is to slow folding and allow abnormal posttranslational modifications to occur. In order to overcome the difficulty of studying collagen itself, the effect of mutations relevant to disease is explored through the study of highly simplified repeating peptides.

This allows detailed biophysical investigations to be carried out, including "real-time" NMR experiments, in order to probe the nature of the folding steps at the level of individual residues (Baum and Brodsky, 1997). Interestingly, this example illustrates both the increasingly well-established link between protein misfolding and human disease (Bychkova and Ptitsyn, 1995), and the power of properly applied structural methods in probing its molecular origins.

I would like to emphasize that most of fibrous proteins (such as silk fibroin, collagen, etc.) are structurally simple owing to the periodicity of their primary structure and, hence, to their large secondary structures.

However, one more addition would not be out of place here.

Proteins forming huge aggregates without any distinct inner structure are also often classified as fibrous proteins. They form a chemically linked elastic matrix in which other more structural proteins are immersed.

Elastin is a typical matrix protein (Stepanov, 1996). It plays an important role in building artery and lung walls, etc. Its long chain is rather hydrophobic and consists of short residue repeats of several types. The resultant product resembles rubber: each elastin chain forms a disordered coil, and together these chains form a net linked by enzyme-modified lysines, four per knot. I cannot but mention that the disturbed function of lysine-modifying enzymes causes the loss of elasticity of vessel walls, and sometimes even a rupture of the aorta.

Note: Sometimes fibrils formed by globules (eg, actin) are regarded as fibrous proteins. We will not consider them here—globular proteins are to be discussed later on.

Instead, I would like to make an additional note on proteins which are usually considered as fibrous ones. It has been relatively recently observed that many "normal" water-soluble proteins (as to membrane proteins, no such data are available yet) are capable of reorganizing so that to form the so-called amyloid fibrils, which are known to be related to human diseases (see Chiti et al., 2003, and references therein). Lysozyme, myoglobin, etc., and some of their mutants, but primarily the notorious prions belong to this group of proteins (see Prusiner, 2012, and references therein).

Amyloid fibrils are insoluble protein aggregates sharing specific structural traits. Though long known, these fibrils were hardly associated with proteins, until it was noticed that prions, these "infectious proteins," form amyloid-like fibrils in the infected cells. Later, it became clear that similar amyloids could also be formed by proteins of a most common nature that seemingly differ from prions only by lack of the ability to transmit the disease from one organism to another (see Afanasieva et al., 2011); however, there are recent reports on infectious nature of the peptide Aβ causing Alzheimer's disease (Nussbaum et al., 2012).

Amyloid fibrils are polymorphic. So far, studied ones had the individual protein chains folded as a double- or multilayer sandwich with, as a rule, the parallel β-structure (Lührs et al., 2005; Nelson et al., 2005), though some polypeptide chains were observed to form no compact globules at all, Fig. 11.9D (Lu et al., 2013) or to display an antiparallel β-structure (Qiang

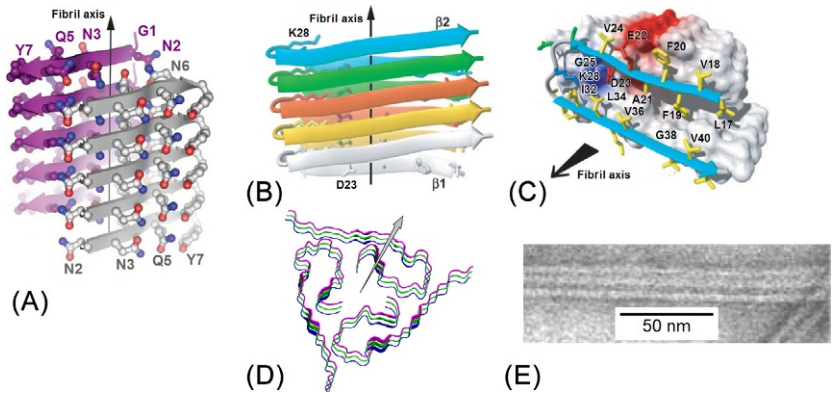


FIG. 11.9 (A) A portion of the X-ray crystallography-solved structure of an amyloid fibril formed by a short (seven residues) fragment of the yeast protein Sup35 (causing a yeast disease similar to a disease of mammals caused by prions; adapted from Nelson et al. (2005), with some simplifications). Note that the "hydrophobic core" of this fibril is composed by *polar* (more so than water) amino acids Asn (N) and Gln (Q), the side groups of which are H-bonded. (B and C) Views from different sides on the structure (solved by a less direct NMR technique) of an amyloid fibril formed by the 42-residue peptide Aβ(1–42), a part of Alzheimer's disease-causing protein; adapted from Lührs et al. (2005), with some simplifications. In the Aβ(1–42) peptide, the fragment 1–17 is disordered (that is why in the figure, only as much as Leu17 is seen), while the rest of it belongs to the protofilament spine whose hydrophobic core comprises mostly "normal" nonpolar amino acids, although the ion pair Asp23-Lys28 is present there as well. (D) The 3D structure of three layers of an amyloid fibril formed by the Aβ(1–40) peptide (model no. 1 from the PDB; Bernstein et al., 1977) structure 2M4J solved by NMR (Lu et al., 2013). (E) Electron microscopy image of a fragment of the amyloid fibril formed by human insulin after 24 h of incubation; the fibril diameter is ≈3−4 nm. *(Adapted from Selivanova, O.M., Suvorina, M.Y., Dovidchenko, N.V., Eliseeva, I.A., Surin, A.K., Finkelstein, A.V., Schmatchenko, V.V., Galzitskaya, O.V., 2014. How to determine the size of folding nuclei of protofibrils from the concentration dependence of the rate and lag-time of aggregation. II. Experimental application for insulin and Lys-Pro insulin: aggregation morphology, kinetics and sizes of nuclei. J. Phys. Chem. B 118, 1198–1206, with some simplifications.)*

et al., 2012). β-Strands are perpendicular to the fibril axis, while interstrand H-bonds are parallel to it (Fig. 11.9A–C). The β-sandwiches are tightly packed as protofilaments, tubular of tape-shaped depending on the protein nature and conditions; the protofilaments in turn form long amyloid fibrils (Fig. 11.9E).

The β-structure usually comprises not the entire protein chain but only a part of it, while the remainder may either keep a globular form or be unfolded and remains off the protofilament spine. Presumably, in amyloid fibrils, a unique spatial structure is typical of solely the sequences involved in β-structural protofilament spine.

Typically, the amyloidogenic sites are rich in amino acids characteristic for the β-structure (see Fig. 11.9C), but sometimes (see Fig. 11.9A) they are composed of Gln- and Asn-rich oligopeptide repeats (Parham et al., 2001).

Amyloid formation includes rearrangement of the protein chain structure, which is not just protein folding (or rather, misfolding) but an aggregation process implying both chain refolding and sticking together many protein molecules. It is known to frequently happen very slowly.

This slowness may underlie the extremely long incubation periods of prion infections like scrapie or mad cow disease. We have already met a simple, "peaceful," and well-studied example of this kind—I mean the structure formation, which is very slow due to a rearrangement. The water-soluble polypeptide poly(Lys), at $pH > 10$ and 20–50°C, undergoes a very fast coil to α-helix transition followed by a very slow conversion of α-helices into the β-structure. The latter can be accompanied by association and can take hours or weeks; we have already discussed it when considering the kinetics of β-structure formation.

Amyloid formation is sometimes a rather complicated process. Some amyloid fibrils assemble through a series of conformational transitions, including even transitions between antiparallel and parallel β-structure (Liang et al., 2014). Amyloid aggregation commonly includes formation of the β-structure and usually has a lag-period, that is, a delay before the beginning of observable fibril growth (Fig. 11.10); this lag-period often is very pronounced.

This lag-period is a result of a very slow spontaneous initiation of fibrils and their subsequent exponential growth and multiplication (due to fragmentation, branching, or "growth from the surface") (Xue et al., 2008; Dovidchenko et al., 2014). In the case of an extremely slow spontaneous initiation, fibril formation can be observed only after introduction of "seeds" (pieces of disrupted mature fibrils) into solution of a potentially amyloidogenic protein (Harper and Lansbury, 1997); when the amyloid formation is initiated by seeding, the lag-period is usually absent.

There are, however, good reasons to believe that the organism is not so much harmed by large mature amyloid fibrils themselves as by short active oligomers

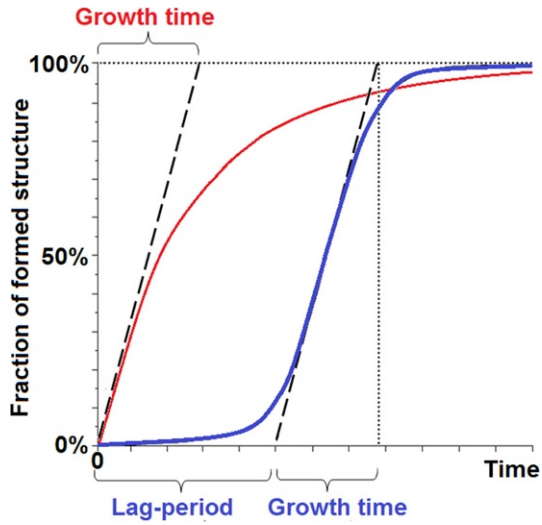


FIG. 11.10 A scheme showing kinetics of unseeded amyloid formation (blue line) where the lag-period is sometimes much shorter and sometimes much longer than the growth time, and kinetics of seeded amyloid formation (red line) where the lag-period is often absent. This kinetics looks rather similar to that of protein folding, which, as we will see soon, is commonly free of any lag-period.

emerging at the beginning of their formation (Ferreira et al., 2007). Anyway, the process and/or the result of amyloid fibril formation underlie serious diseases which may have an extremely long incubation period (eg, kuru, mad cow disease, Alzheimer's disease, etc.). These and other "conformation-related" diseases had been long predicted by Bychkova and Ptitsyn (1995); they now are the "hot spots" in molecular biology research.

REFERENCES

Afanasieva, E.G., Kushnirov, V.V., Ter-Avanesyan, M.D., 2011. Interspecies transmission of prions. Biochemistry (Mosc) 76, 1375–1384.

Alexandrov, V.Ya., 1965. On the biological significance of the correlation between the level of thermoresistance of proteins and the environmental temperature of the species. Usp. Sovr. Biol. 60, 28–44 (in Russian).

Baum, J., Brodsky, B., 1997. Real-time NMR investigations of triple-helix folding and collagen folding diseases. Fold. Des. 2, R53–R60.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M., 1977. The protein bank. A computer-based archival file for macromolecular structures. Eur. J. Biochem. 80, 319–324. modern version: http://www.wwpdb.org/.

Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second ed. Garland Publishing Inc., New York, London (Chapter 3).

Bychkova, V.E., Ptitsyn, O.B., 1995. Folding intermediates are involved in genetic diseases? FEBS Lett. 359, 6–8.

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry: Part 1. W.H. Freeman & Co., New York (Chapter 2).

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., Dobson, C.M., 2003. Rationalisation of mutational effects on protein aggregation rates. Nature 424, 805–808.

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

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W.H. Freeman & Co., New York, NY (Chapter 5).

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

Dickerson, R.E., 1964. X-ray analysis of protein structure. In: Neurath, H. (Ed.), second ed. In: The Proteins, vol. 2. Academic Press, New York, NY, pp. 603–778.

Dovidchenko, N.V., Finkelstein, A.V., Galzitskaya, O.V., 2014. How to determine the size of folding nuclei of protofibrils from the concentration dependence of the rate and lag-time of aggregation. I. Modeling the amyloid photofibril formation. J. Phys. Chem. B 118, 1189–1197.

Ferreira, S.T., Vieira, M.N., De Felice, F.G., 2007. Soluble protein oligomers as emerging toxins in Alzheimer's and other amyloid diseases. IUBMB Life 59, 332–345.

Harper, J.D., Lansbury Jr., P.T., 1997. Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time dependent solubility of amyloid proteins. Annu. Rev. Biochem. 66, 385–407.

Kreplak, L., Doucet, J., Dumas, P., Briki, F., 2004. New aspects of the α-helix to β-sheet transition in stretched hard α-keratin fibers. Biophys. J. 87, 640–647.

Lehninger, A.L., Nelson, D.L., Cox, M.M., 1993. Principles of Biochemistry, second ed. Worth Publishers, New York (Chapter 7).

Liang, C., Ni, R., Smith, J.E., Childres, W.S., Mehta, A.K., Lynn, D.G., 2014. Kinetic intermediates in amyloid assembly. J. Am. Chem. Soc. 136, 15146–15149.

Lu, J.X., Qiang, W., Yau, W.M., Schwieters, C.D., Meredith, S.C., Tycko, R., 2013. Molecular structure of beta-amyloid fibrils in Alzheimer's disease brain tissue. Cell 154, 1257–1268.

Lührs, T., Ritter, C., Adrian, M., Riek-Loher, D., Bohrmann, B., Döbeli, H., Schubert, D., Riek, R., 2005. 3D structure of Alzheimer's amyloid-beta (1–42) fibrils. Proc. Natl. Acad. Sci. U. S. A. 102, 17342–17347.

Nelson, R., Sawaya, M.R., Balbirnie, M., Madsen, A.Ø., Riekel, C., Grothe, R., Eisenberg, D., 2005. Structure of the cross-beta spine of amyloid-like fibrils. Nature 435, 773–778.

Nussbaum, J.M., Schilling, S., Cynis, H., Silva, A., Swanson, E., Wangsanut, T., Tayler, K., Wiltgen, B., Hatami, A., Rönicke, R., Reymann, K., Hutter-Paier, B., Alexandru, A., Jagla, W., Graubner, S., Glabe, C.G., Demuth, H.U., Bloom, G.S., 2012. Prion-like behaviour and tau-dependent cytotoxicity of pyroglutamylated amyloid-β. Nature 485, 651–655.

Parham, S.N., Resende, C.G., Tuite, M.F., 2001. Oligopeptide repeats in the yeast protein Sup 35 p stabilize intermolecular prion interactions. EMBO J. 20, 2111–2119.

Prusiner, S.B., 2012. Cell biology. A unifying role for prions in neurodegenerative diseases. Science 336, 1511–1513.

Qiang, W., Yau, W.-M., Luo, Y., Mattson, M.P., Tycko, R., 2012. Antiparallel β-sheet architecture in Iowa-mutant β-amyloid fibrils. Proc. Natl. Acad. Sci. U. S. A. 109, 4443–4448.

Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York/Heidelberg/Berlin (Chapters 4 and 5).

Stepanov, V.M., 1996. Molecular Biology: Protein Structure and Function. Vysshaya Shkola, Moscow (Chapter 5, in Russian).

Stryer, L., 1995. Biochemistry, fourth ed. W.H. Freeman & Co., New York (Chapter 9).

Traub, W., Piez, K.A., 1971. The chemistry and structure of collagen. Adv. Protein Chem. 25, 243–352.

Volkenstein, M.V., 1977. Molecular Biophysics. Academic Press, New York/London (Chapter 4).

Xue, W.F., Homans, S.W., Radford, S.E., 2008. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. Proc. Natl. Acad. Sci. U. S. A. 105, 8926–8931.

This page intentionally left blank

Lecture 12

Let us now focus on membrane proteins. As concerns their specific transmembrane parts "living" in hydrophobic environment, these proteins are almost as simple as fibrous proteins (Creighton, 1993; Stryer, 1995; Branden and Tooze, 1991; Nelson and Cox, 2012).

Membranes are films of lipids (fat) and protein molecules (Fig. 12.1). They envelop cells and closed volumes within them (the so-called "compartments"). The peculiar role of membrane proteins (they amount to half of the membrane weight) is to act as a power plant dam wall in the cell (this we will consider later) and to provide transmembrane transport of various molecules and signals—this we will consider now.

In transmitting a signal, the lipids of a membrane work as a kind of "insulator" while its proteins (or rather, as we will see later, protein channels) act as "conductors." These conductors are specific, each ensuring transmembrane transport of molecules of a particular kind or signals from particular molecules (by a slight change in the protein's conformation).

True membrane proteins reside within the membrane where there is virtually no water. The intramembrane parts have a regular secondary structure, and their size is determined by the membrane thickness (Fig. 12.1).

Let us consider the structure of membrane proteins using several examples.

As a matter of fact, structures of membrane proteins so far constitute $<1\%$ of protein structures collected in the Protein Data Bank (Berman et al., 2012), although membrane proteins constitute about one-third of all proteins (Neumann et al., 2012). This is due to their poor solubility in water (detergents have to be used, etc.) and difficulties of crystallization caused by their tendency to disordered association (Huber, 1989).

Fig. 12.2 gives the structure of bacteriorhodopsin which pumps protons across the membrane. This structure was originally determined from many high-resolution electron microphotographs (Amos et al., 1982; Henderson et al., 1990), because it was too difficult to obtain good 3D crystals of bacteriorhodopsin. Therefore, its X-ray structure (which appeared to be very much the same) was reported much later (Pebay-Peyroula et al., 1997).

As we can see, the transmembrane portion of bacteriorhodopsin comprises seven regular α-helices that form a membrane-spanning bundle slightly tilted with respect to the plane of the membrane, while the single β-hairpin and all irregular segments (connecting loops) protrude from the membrane.

The highly regular arrangement of the transmembrane chain backbone is only natural. Each H-bond is expensive in the fatty, waterless environment. Therefore, the protein chain has to adopt a structure with fully accomplished hydrogen bonding, that is, either the α-helix or the β-cylinder (see later).

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00012-0

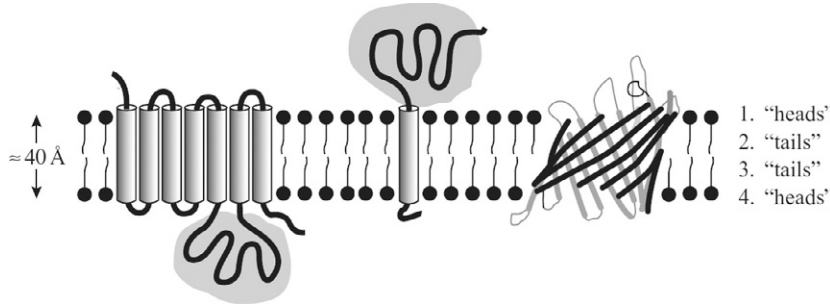


FIG. 12.1 Membrane-embedded proteins. Extramembrane domains are shown, very schematically, in gray. Protein sequences within the membrane are virtually free of irregular segments. In eukaryotes, chain portions projecting from the membrane out of the cell are strongly glycosylated and therefore more hydrophilic. 1. Polar "heads" of lipids of one membrane layer. 2. Nonpolar "tails" of lipids of the same membrane layer. 3. Non-polar "tails" of lipids of second membrane layer. 4. Polar "heads" of lipids of second membrane layer. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

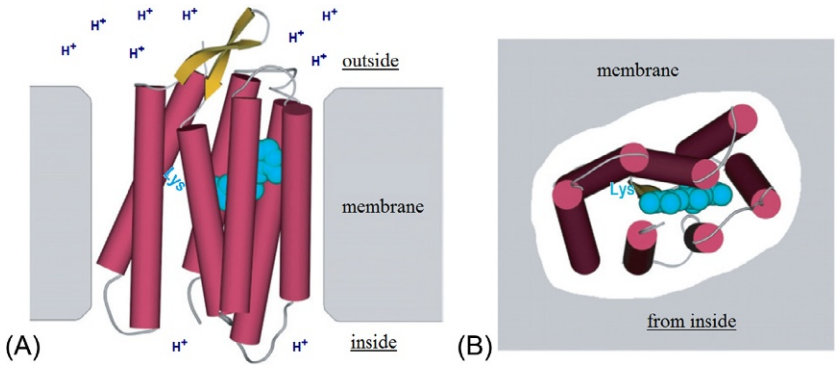


FIG. 12.2 Membrane-embedded bacteriorhodopsin: (A) as viewed along the membrane; (B) as viewed from the bottom (from the cytoplasm). Its seven helices are shown as cylinders. The connecting loops are also shown together with the bound retinal molecule (light-blue). The lipid layer is schematic. H^+ concentration presented in (A) is greatly exaggerated: the distance between the ions shown in (A) is ~1 nm, while at pH7 this distance is ~250 nm. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

The hydrophobic groups positioned on the bacteriorhodopsin α-helices are turned rather "outwards," towards lipids that are also hydrophobic, while the few polar groups face the interior and form a very narrow proton-conducting channel. The proton flow is mediated by a cofactor, which is the retinal molecule chemically bound to Lys of helix G inside the bundle of helices. Retinal blocks the central channel of bacteriorhodopsin.

Bacteriorhodopsin is a light-driven proton pump. Having accepted a photon, retinal changes its form, from *trans* to *cis* (Fig. 12.3), that is, bends by rotation about the bond $C_{13}=C_{14}$ (with a simultaneous slight change of the protein body conformation) and moves a proton (H^+) from one (inside) end of the seven-helix bundle to the other, extracellular end. Then retinal regains its previous *trans*-shape, but this time without a proton (Branden and Tooze, 1991), and then receives H^+ from the cytoplasm. In this way the H^+ pump works and transports H^+ from the cytoplasm (where there are few H^+) to the outside (where there are a lot of them). That is, the photon-induced H^+ transport goes *against* the H^+ concentration gradient.

The point is that retinal-bound N-atom of Lys (Fig. 12.3) binds H^+ very strongly when retinal is in the *trans* form; and the bacteriorhodopsin construction is such that the retinal having its ground-state *trans* form turns its H^+-binding site towards the inner side of the membrane, and this strong binding site only can bind H^+ from this (cytoplasm) side, although H^+ concentration here is low. But when retinal has absorbed a photon, obtaining its activated *cis* form and bent, it binds H^+ very loosely, and simultaneously, the H^+-binding site faces the outer side of the membrane. Thus, the loose binding site releases H^+, which only can go to the outer side of the membrane, although H^+ concentration there is high.

It is interesting that the same retinal works differently in rhodopsin (a protein quite similar to bacteriorhodopsin), though it is bound to Lys in both of these proteins. In rhodopsin, the retinal ground state is *cis* (rather than *trans*, as in bacteriorhodopsin); its photoactivated state is *trans* (rather than *cis*); and the *cis-trans* transition in rhodopsin-localized retinal results from a turn around the $C_{11}=C_{12}$ bond (rather than around the $C_{13}=C_{14}$ one, as in bacteriorhodopsin). This demonstrates how the protein environment can change the behaviour of a co-factor.

Bacteriorhodopsin's central channel is narrow. Similar (but usually wider) pores arranged like a hollow bundle of helices can be formed in other cases from separate α-helical transmembrane peptides (Branden and Tooze, 1991).

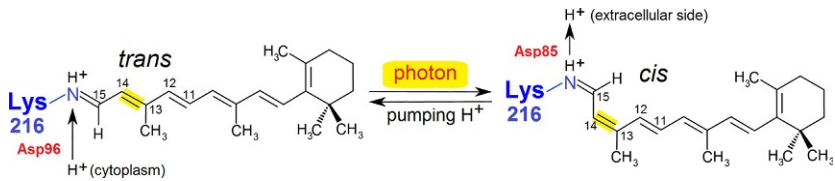


FIG. 12.3 Photoisomerization of retinal in bacteriorhodopsin. Retinal is bound to the N atom of the bacteriorhodopsin's Lys; this N atom (shown in blue) accepts and releases H+. Two other groups of bacteriorhodopsin that facilitate binding (Asp96) and release (Asp85) of H^+ are also shown. *(Adapted from https://en.wikipedia.org/wiki/Bacteriorhodopsin; https://en.wikipedia.org/wiki/Retinal.)*

By the way, helical bundles are also abundant in quite different membrane proteins. These do not transport molecules across the membrane but conduct signals.

I am now speaking about receptors, and specifically about hormone receptors (Fig. 12.4). There are two principal signal transduction pathways involving the so-called G protein-coupled receptors: the phosphatidylinositol signal pathway and the cyclic adenosine monophosphate (cAMP) signal pathway. Let us briefly consider one of many G protein-coupled receptors connected with the cAMP signal pathway. I will use β_2-adrenergic G protein-coupled receptor as an example; like many receptors of this kind, it contains seven helices and looks very much like bacteriorhodopsin (Cherezov et al., 2007).

Having bound a hormone, this receptor somehow changes the conformation of its transmembrane helical bundle thereby "announcing" the hormone's arrival (for details, see Cherezov et al., 2007; Trzaskowski et al., 2012), but many of them are not completely clear. This signal causes the α-subunit of the receptor-associated *G-protein* (*Guanine-binding protein*) to release its own guanosine diphosphate (GDP) molecule and take up a guanosine triphosphate (GTP) molecule from the surrounding cytosol. Then this α-subunit leaves both its fellow subunits and the receptor, thus providing an opportunity for another GDP-loaded G-protein α-subunit to contact the receptor and then, in its turn, to leave it having exchanged its GDP for GTP (Fig. 12.4).

The α-subunit of G-protein can cleave GTP (guanidine triphosphate) and convert it to GDP (guanidine diphosphate) and phosphate (P) but, importantly, the process is slow. That is, it is has been tailored by the Nature as a "very bad" enzyme; the reason for the slow enzyme action will be considered in the last part of this course. Meanwhile, together with GTP, the α-subunit (with its "tail" buried inside the membrane) drifts along the membrane, reaches adenylate

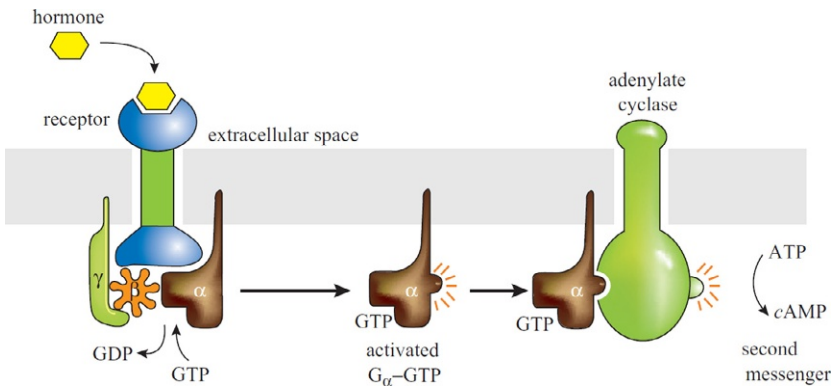


FIG. 12.4 Activation of adenylate cyclase by G-protein, which, in turn, is activated by the hormone-binding transmembrane receptor. *(Reproduced from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12). Reproduced by permission of Garland Science/Taylor & Francis Group LLC.)*

cyclase and binds to it; as a result, adenylate cyclase starts functioning and converts many molecules of adenosine triphosphate into cAMP. This initiates a physiological reaction of the cell in response to hormone binding. But the α-subunit's impact upon adenylate cyclase eventually ends when α-subunit-bound GTP turns into GDP. Then, with this GDP, the α-subunit drifts along the membrane and eventually comes back to the host receptor. If the latter is still bound to hormone, the cycle repeats; if not, it is over (Sprang, 1997; Branden and Tooze, 1991).

Hence, the signal of a molecule of hormone is enhanced many-fold, but its duration is finite. This outlines a peculiarity of all G-proteins (not only those activating adenylate cyclase): they function until GTP is cleaved, and a slow GTP cleavage serves as a peculiar timer.

G protein-coupled receptors are found only in eukaryotes (King et al., 2003), where they play an extremely important role in various types of signal transduction. They are involved in many diseases and are also the target of approximately 40% of all modern medicinal drugs (Filmore, 2004; Overington et al., 2006). So, no wonder that many Nobel Prizes have been awarded for various aspects of G protein-mediated signaling.

Now we will consider porin (Fig. 12.5), another transmembrane protein. Its structure is highly regular too and looks like a wide cylinder built up from β-structures. Note that here the β-sheet forms a *closed* β-cylinder, thus avoiding the "free" H-bond donors and acceptors typical of edges of a planar β-sheet. The cylinder comprises 16 very long β-segments, and the diameter of a pore in its center is about 10 Å. The side-groups of polar residues pertaining to the β-strands face the pore, while nonpolar residues alternating with them in the strand face the membrane.

Porin is responsible for the transport of polar molecules, but it is not very selective (Branden and Tooze, 1991).

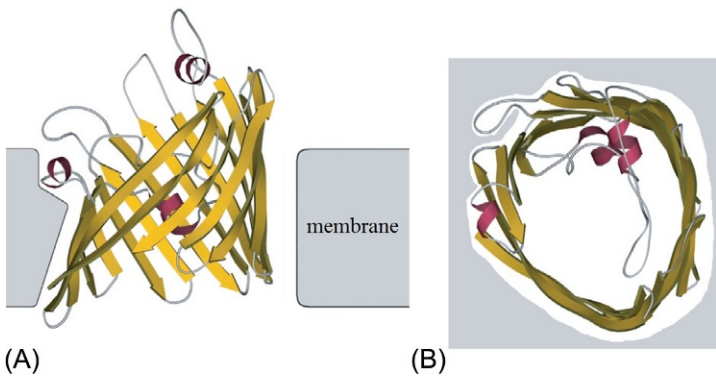


FIG. 12.5 Porin, as viewed (A) along and (B) across the membrane plane. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

The transport selectivity, that is, specificity of function, of membrane proteins is, of course, dependent on the pore diameter and size of the molecule that tries to pass through it: too large molecules cannot penetrate into it (but the penetration is also difficult for molecules that are a little too small, because (Finkelstein et al., 2006) they are separated from the channel walls by a vacuum gap that is too (narrow to accommodate a water molecule).

The selectivity also is strongly determined by the fact that separate polar groups, not to mention charged ones, can hardly penetrate inside the membrane by themselves.

As you may remember, the free energy of a charge q in the medium with permittivity ε is equal to $+q^2/2\varepsilon r$, where r is the van der Waals radius of the charge. It can be easily estimated that, with q equal to the electron charge and $r = 1.5$ Å (the typical radius of a singly charged ion), the value of $+q^2/2\varepsilon r$ is close to 1.5 kcal mol^{-1} at $\varepsilon = 80$ (ie, in water), while at $\varepsilon_{\text{membr}} = 3$ (ie inside a "purely lipid" membrane), this value will amount to nearly 37 kcal mol^{-1}. In total, an increase in the free energy $\triangle F$ of +35 kcal mol^{-1} results. According to Boltzmann, the probability of accumulation of such free energy is $\exp(-\triangle F/kT) = \exp(-35/0.6) = 10^{-25}$. This means that only one in 10^{25} ion attacks on the membrane will be successful. Given the attack time is no less than 10^{-13} s (as you know, this is the thermal fluctuation time), for an ion passing through a purely lipid membrane would take at least $\sim 10^{12}$ s, that is about 10,000 years. Thus, a purely lipid membrane appears to be virtually impermeable for ions.

It is quite another matter if the membrane includes protein with a more or less broad water-filled channel where ions at least partly enjoy the high permittivity of water, although to some extent restricted by the surrounding membrane. Roughly (Finkelstein et al., 2006), the membrane-caused increase in the ion's free energy amounts to about $+q^2/[(\varepsilon_{\text{membr}}\ \varepsilon_{\text{water}})^{1/2}\ R]$, where R is the channel radius, $\varepsilon_{\text{membr}} = 3$ and $\varepsilon_{\text{water}} = 80$. It is easy to calculate that with $R \approx 1.5$ Å, it takes an ion a fraction of a second to pass through the channel, and with $R \approx 3$ Å the time is a tiny fraction of a millisecond.

The channel sites that can attract (or repulse) the ion and thereby reduce (or increase) the barrier to be overcome regulate the selectivity of ion transfer across the membrane. For example, the presence of a positive charge near the channel accelerates transport of negatively charged ions and strongly hampers transport of positively charged ions (Fig. 12.6). In the case of a negatively charged channel, the transport of positive ions is accelerated while that of negative ions is hampered. This effect (and, of course, the pore diameter) underlies the selectivity of membrane proteins responsible for the transport of molecules.

Fig. 12.6 is merely a scheme, but Fig. 12.7 (taken from Long et al., 2007) shows the key part of a transmembrane pore, the "filter," in the potassium channel. Such channels are found in most cell types; they control a wide variety of cell functions and are targets for many drugs (Jessell et al., 2000; Hille, 2001;

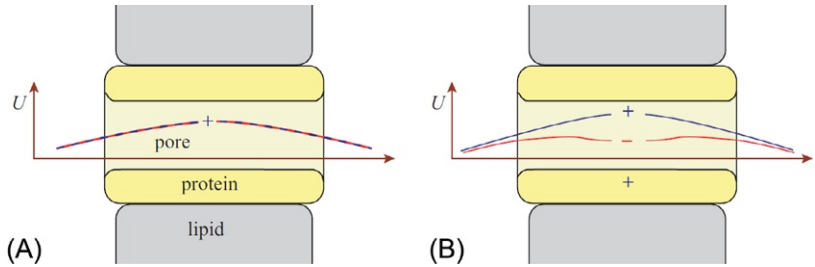


FIG. 12.6 Very schematic diagram of a transmembrane pore (here the membrane is vertical and the pore is horizontal) and the electrostatic free energy U for positively (- - -+- - -) and negatively (——– ——) charged ions: (A) with no charge on the inner pore surface; (B) with a positive charge close to the pore.

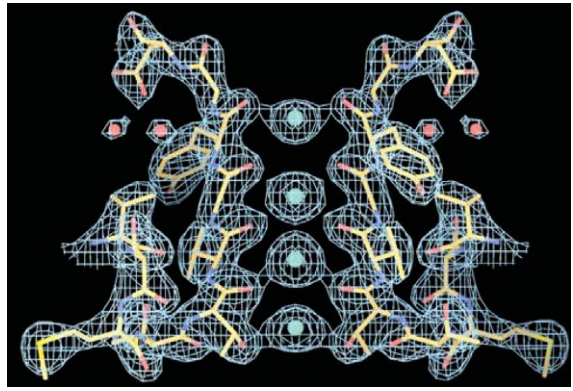


FIG. 12.7 The bottleneck of the potassium channel: the K^+ selectivity filter. For clarity, the parts of only two out of four subunits surrounding the channel are shown (yellow/red sticks surrounded by electron density contours). K^+ (blue spheres) surrounded by electron density contours) and water molecules (small red spheres surrounded by electron density contours) are also shown. *(Adapted from Long, S.B., Tao, X., Campbell, E.B., MacKinnon, R., 2007. Atomic structure of a voltage-dependent K^+ channel in a lipid membrane-like environment. Nature 450, 376–382, with permission.)*

Rang, 2000).The membrane voltage-dependent channel is shown here in its "open," K^+-transmitting state. One can see that the pore diameter exactly fits the potassium (K^+) ion size, and that walls of the pore are paved with carbonyl oxygens (red short sticks in Fig. 12.7), which are strongly electro-negative and attractive for cations (ie, "plus" charges). The shown intramembrane filter selects and transmits K^+ (and does not transmit smaller Na^+ (Armstrong, 1998); the explanation is given above, while the voltage-sensitive gating (controlling opening and closing of the channel) is performed by a significant displacement of the extramembrane domains of the same protein.

Now let us focus on the photosynthetic reaction center (Fig. 12.8) (Deisenhofer et al., 1985; Huber, 1989; Branden and Tooze, 1991). Its function is to ensure the transport of light-released electrons from one side of the membrane to the other, thereby creating the transmembrane potential that underlies photosynthesis.

The photosynthetic reaction center comprises cytochrome with four hemes (actually, this protein is not a membrane protein: it is outside the membrane in the periplasmic space) and three membrane subunits, L, M, and H (though the transmembrane part of the last is represented by one α-helix only). Subunits L and M are very much alike.

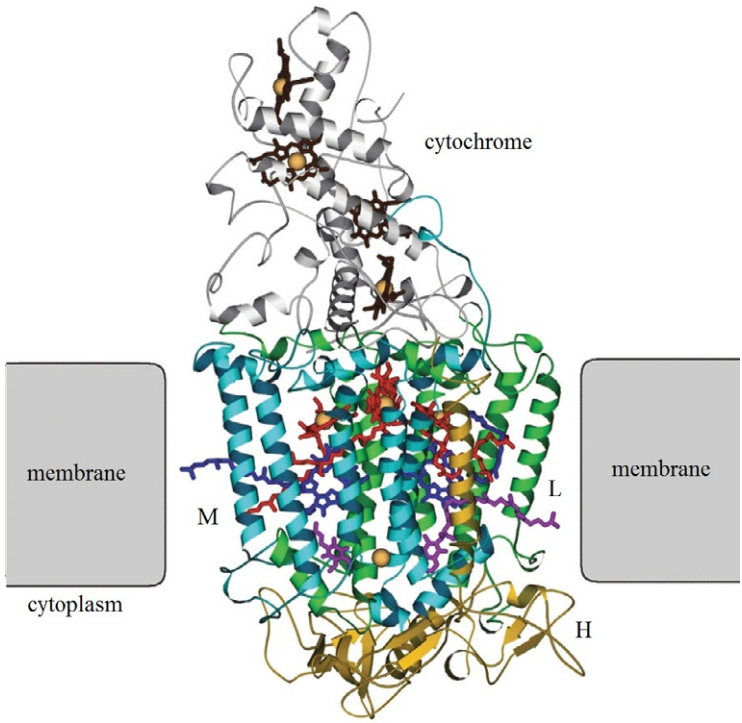


FIG. 12.8 A photosynthetic reaction center. The membrane is shown schematically. The transmembrane subunit M appears in light-blue, L in green and H (with its single transmembrane helix) in yellow, while cytochrome is gray. Notice how distinctly more regular the transmembrane chain portions are compared with those outside the membrane. The subunits L and M bind photosynthetic pigments, chlorophylls (shown in red with a yellow spot for the magnesium ion) and pheophytins (shown in dark-blue). Each has a long hydrophobic tail projecting from the protein into the membrane. The subunits L and M also bind the two quinones, shown in violet. Cytochrome positioned outside the membrane binds four hemes (grayish-black with yellow spots for iron ions). All cofactors are shown as wire models; see also Fig. 12.9. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

All transmembrane parts are α-helical. As usual, they are long (equal to the membrane thickness) and regular. There are no irregular loops inside the membrane. The outer chain portions are considerably less regular and contain many loops; in fact, their fold is the same as that of "ordinary" water-soluble proteins, which we will discuss later.

Notice the many rather small cyclic molecules, pigments, embedded in this protein: these form the "conductors," that is, the pathways for the electron flow (the flow of electrons can be followed by the changing electron spectra of pigments during electron transport). The polypeptide only serves as a "shaping insulator."

First, a light quantum (either originating directly from sunlight or transferred as excitation energy via a light-harvesting antenna system, see later) displaces an electron from the "special pair" of chlorophylls (see the schematic diagram, Fig. 12.9), where it has low energy, to the chain of pigments where it has an elevated energy. Having passed (within a few picoseconds) through the "accessory" chlorophyll B_A, this electron then instantly, within a picosecond, joins to pheophytin P_A (note: P_A and not P_B), and about 200 ps later, it arrives at quinone Q_A. Then it spends about a fraction of millisecond to reach Q_B (Branden and Tooze, 1991; Leonova et al., 2011). We still do not know why the electron prefers this roundabout way to Q_B.

An electron coming from the cytochrome heme replaces the electron released from the special pair of chlorophylls. This completes the first half-cycle of the reaction.

The other similar half-cycle brings another electron to Q_B making it doubly charged and therefore capable of leaving the membrane more easily to participate in further photosynthetic events.

Thus, the photosynthetic reaction center performs electron transport from the upper (in Figs. 12.8 and 12.9) to the lower compartment, that is, against the apparent difference in electric potential between the compartments. The efficiency is about 50% (in other words, 50% of the captured light is converted into the energy of separated charges, which is not bad at all).

The following three important physical aspects must be emphasized:

1. If a special pair of bacteriochlorophylls were to absorb light alone, it would use only a small portion of the light flux. However, the photosynthetic reaction center is surrounded by a light-harvesting antenna. The antenna is not directly involved in photochemical reactions, but is designed to absorb light and transfer excitations toward the photosynthetic reaction center (Liu et al., 2004; Barros and Kuhlbrandt, 2009) consisting (in bacteria) of one "large" complex LH1 (probably embracing the photosynthetic reaction center) and 8–10 "small" complexes LH2 located around (see Branden and Tooze, 1991; Cogdell et al., 2003, and references therein). The LH1 complex is composed of 32 transmembrane α-helices collected in a hollow cylinder, and 32 bacteriochlorophylls seating between these helices and providing

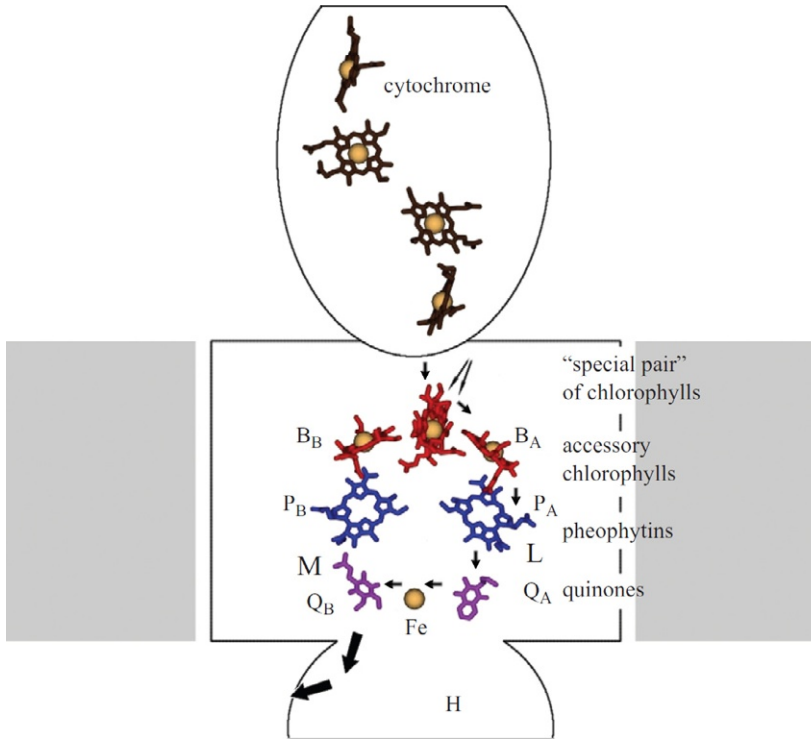


FIG. 12.9 Schematic arrangement of the photosynthetic pigments in the reaction center. The reaction center orientation is the same as in Fig. 12.8. The pseudo-twofold symmetry axis of the L and M subunits passes through the "special pair" of chlorophylls and the Fe ion. The pigments' long "tails" are omitted as unnecessary details. Electron transfer proceeds preferentially along the branch associated with the L subunit (on the right of the figure). The electron pathway is shown as small arrows. The large arrow indicates release of the quinone with two consecutively accepted electrons. The left chain is not used. Presumably, it was used in the past, but in present-day reaction centers it is an appendix-like relic. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

their contacts. Each "small" LH2 complex consists of 18 transmembrane α-helices forming a hollow cylinder, and 27 bacteriochlorophylls seating between these helices and providing their contacts

Thus, a special pair of bacteriochlorophylls of the photosynthetic reaction center is surrounded by about 300 other bacteriochlorophylls. Upon reaching any bacteriochlorophyll of this huge system, a photon excites the collective resonance energy transfer embracing entire bacteriochlorophyll system, and this excitation persists in the system until it knocks out an electron from the special pair of bacteriochlorophylls.

2. All chlorophylls (Fig. 12.10), pheophytins, quinones, and other pigments contain partial double (p-electron) bonds of the $\cdots\text{–}C=C\text{–}C=\cdots$ type. In other words, owing to Pauling resonance ($\cdots C=C\text{–}\cdots \Leftrightarrow \cdots C\text{–}C=\cdots$), each

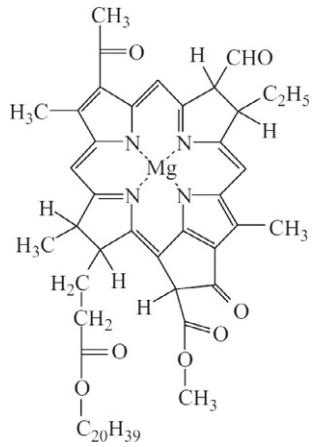


FIG. 12.10 A molecule of bacteriochlorophyll. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).)*

FIG. 12.11 Schematic representation of tunneling. The profile of the electron potential energy U is shown as a solid line. The dashed line shows the level of total (potential + kinetic) energy of the electron. The bell-shaped line (with hatching below) represents the density of the electron cloud ρ. Initially, the electron stays in the left well (as shown in the figure) with the edge of its cloud (although at an extremely low density) reaching the right well, where the electron can move with time if the energy proves to be lower there. Given a well depth of a few electron Volts or about 100 kcal mol^{-1} (the typical energy required for molecular ionization), the typical distance at which the density of the electron cloud is reduced by an order of magnitude is about 1 Å.

of these molecules is covered by a common electron cloud, and on such a molecule electrons move as on a piece of metal. This provides a potential well where electrons are "delocalized," that is, where they can shift by distances greater than the atom diameter. (Note that it is electron delocalization that underlies the typical pigment colors: an electron localized in a separate covalent bond is excited by short-wave UV light, whereas a delocalized electron is excited by "ordinary" visible light of a longer wavelength.)

3. Electron transfer from one "piece of metal" (pigment) to another requires no direct contact of these pigments. It is performed by *quantum tunneling* (see Fig. 12.11).

The main point about quantum tunneling (also called the "sub-barrier" transition because it is as if the electron passes *under* the energy barrier), is that in accordance with quantum mechanics, an electron (like any other particle,

and especially a light particle) "projects" slightly beyond the potential well in which it resides (Landau and Lifshitz, 1977). The host molecule of the electron (chlorophyll, pheophytin, etc.) serves as the "potential well," that is, the area of low potential energy U.

When outside the "well" (Fig. 12.11), the electron's potential energy is higher than its total (potential + kinetic) energy when in the well. But for the quantum effect, this deficient energy would not let the electron density project a straw beyond the well. However, owing to the quantum effect the electron wave function (or, simply, the density of the electron cloud) extends beyond the potential well, although the value of this density decreases exponentially with growing distance from the well.

The latter point is another manifestation of the same quantum effect that prevents an electron from falling onto the nucleus: although this would decrease the potential energy of the electron, its kinetic energy would increase still more. The thing is that if the distance between the electron and the nucleus ($\triangle x$) tends to zero, the potential energy of the electrostatic interaction of the electron with the nucleus tends to minus infinity as $1/(\triangle x)$, while according to the Heisenberg Uncertainty Principle, the electron's kinetic energy tends (at $\triangle x \to 0$) to plus infinity far more rapidly, as $1/(\triangle x)^2$.

Indeed, according to the Heisenberg principle, uncertainty in speed ($\triangle v$) and uncertainty in coordinate ($\triangle x$) of a particle are related as $m \triangle v \triangle x \sim \hbar$, where $\hbar$ is Planck's constant divided by 2π, and m is the mass of the particle. In other words, the absolute value of the particle's speed (v) in the well of a $\triangle x$ width is about $\hbar/(m\triangle x)$ (with complete uncertainty of the direction in which the particle moves at the given moment). Hence, the kinetic energy of the particle, $E = mv^2/2$, is a value of about $m[\hbar/(m\triangle x)]^2 = (\hbar^2/m)\triangle x^2$.

The same is true for an electron staying in the potential well: provided it does not project a straw beyond the well, its total energy would be higher.

That is why the electron slightly "comes out" of the potential well and its density decreases exponentially, like the electron cloud of an atom. The typical distance at which the cloud density becomes an order of magnitude (10-fold) lower is about 1 Å (which, as we know, is the typical radius of an atom).

Clarification. Quantum-mechanical calculation (see Landau and Lifshitz, 1977) shows that the characteristic distance λ, where electron cloud density decreases by 10 times, is, approximately, $\ln 10 \cdot \hbar / \sqrt{8m \cdot \Delta E}$ (where m is the electron mass and $\Delta E > 0$ is the difference between the barrier energy and the level of electron's energy in the well); a typical value of $\Delta E \sim 5$ electron-volts (≈ 120 kcal mol^{-1}) leads to $\lambda \approx 1$ Å.

Hence, the electron cloud density becomes 1000 times lower at a distance of 3 Å from the "home well." This means that the probability of electron's moving as far as 3 Å off its "home well" during one vibration is about 10^{-3} (for 5 Å it is about 10^{-5}, for 10 Å about 10^{-10}, and so on). When in the "well" (in a pigment molecule), an electron performs $\sim 10^{15}$ vibrations per second (it vibrates at visible light frequencies: this is seen from light absorption spectra of such

molecules). Hence, the typical time of its transition into another "well" (another pigment molecule) 3 Å away is about 10^{-15} s/$10^{-3} = 10^{-12}$ s; for 5 Å it is about 10^{-10} s, for 10 Å about 10^{-5} s and for 15 Å about 1 s. This simple relationship between transition rates and distances is in a qualitative agreement with what we observe in the photosynthetic reaction center.

The following points deserve your special attention:

First. The total distance of the electron transition in the photosynthetic center is about 40 Å. This distance cannot be covered by one tunnel jump (it would take $\sim 10^{-15}$ s/$10^{-40} \sim 10^{25}$ s or $\sim 10^{17}$ years, which is beyond the lifetime of the Universe). However, because of the protein arrangement, this great jump is divided into four small jumps from one electron-attracting pigment to another; as a result, an electron covers the entire 40 Å distance during a fraction of a millisecond.

Second. To prevent an electron's prompt return to the first pigment from the second one, and to promote its further movement to the third pigment, and so on, its total (potential + kinetic) energy must decrease along the pathway; in other words, every step of the electron path must be a descent from a high-energy orbital to a low-energy one. The arrangement of the photosynthetic reaction center is believed to provide such a decrease in electron energy from pigment to pigment.

Third. An electron spends no energy on tunneling (since there is no "friction" here). The energy is decreased by the electron-conformational interaction (Volkenstein, 1979). Specifically, when arriving at the next pigment, an electron faces the pigment's conformation corresponding to the energy minimum *without* the newcomer. In its presence, the energy minimum corresponds to another, slightly deformed, conformation of the pigment (ie, another location of the nuclei of its atoms). When adopting this new conformation, the pigment atoms rub against the surroundings, and the excess energy dissipates. As a result, at each step, the electron changes a high-energy orbital for a low-energy orbital, its energy decreases and appears to be spent on making the tunnel transition "efficient," that is, irreversible.

Fourth. A tunneling (sub-barrier) transition can be distinguished from an ordinary activation mechanism that requires the overcoming of an energy barrier (Frauenfelder, 2010). The rate of tunneling is virtually temperature-independent (and therefore tunneling does not disappear at low temperatures), whereas the rate of an activation transition (proportional to $\exp(-\triangle E^{\#}/k\mathrm{B}T)$, where $\triangle E^{\#}$ is the energy of the activation barrier, and T is temperature) decreases dramatically with decreasing temperature.

In conclusion, let me say a few words on folding of membrane proteins. Some membrane proteins fold spontaneously (Roman and González Flecha, 2014) like water-soluble proteins; this phenomenon we will consider in one of the subsequent lectures. But usually membrane protein folding is not spontaneous; rather, it is assisted by some special cellular mechanisms (Skach, 2009).

- *Inner voice:* Do you mean chaperones?
- *Lecturer:* From my point of view, real chaperones are used just to fold membrane proteins. But for some historical reasons which I do not know, the term "chaperones" is only applied to proteins that somehow, and in any case much less manifestly, assist folding the water-soluble proteins. We will come to chaperones in one of the subsequent lectures.

As far as membrane proteins are concerned, the mentioned "special cellular mechanisms" solve two problems:

1. To establish protein topology by selective peptide transport to the opposite sides of a cellular membrane;
2. To insert, integrate, and fold transmembrane segments within the lipid bilayer.

In eukaryotes, the membrane protein folding usually takes place in the endoplasmic reticulum, coincident with protein synthesis, and is facilitated by the "translating ribosome and the translocon" complex (Fig. 12.12). At its core, this complex is assumed to form a dynamic pathway through which the elongating nascent polypeptide moves as it is delivered. The complex is assumed to function as a protein-folding machine that restricts conformational space by establishing transmembrane topology (maybe with the aid of some factors that bind to specific signals contained in the nascent protein sequence) and yet provides a permissive environment that enables nascent transmembrane domains to efficiently fold.

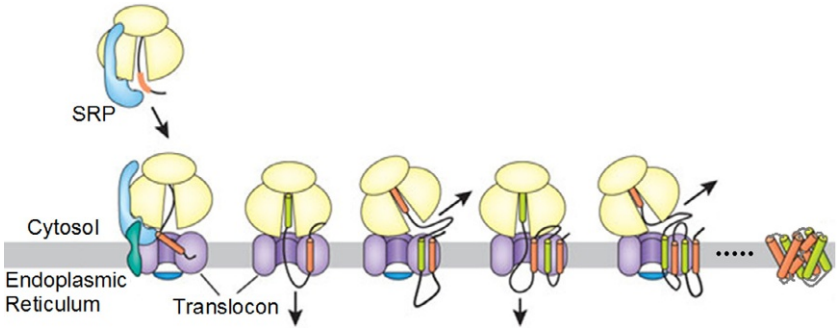


FIG. 12.12 Tentative mechanism of transmembrane integration. For explanations, see the text. SRP is the signal protein recognition particle that binds to the signal sequence of a nascent peptide as it emerges from the ribosome and targets them both to the membrane. Arrows show directions of protrusion of the loops. The translocon (the structure of which remains unknown) is a protein complex that is involved in translocation of the nascent protein chain across the membrane and its folding. *(Adapted from Skach, W.R., 2009. Cellular mechanisms of membrane protein folding. Nat. Struct. Mol. Biol. 16, 606–162.)*

REFERENCES

Amos, L.A., Henderson, R., Unwin, P.N., 1982. Three-dimensional structure determination by electron microscopy of two-dimensional crystals. Prog. Biophys. Mol. Biol. 39, 183–231.

Armstrong, C., 1998. The vision of the pore. Science 280, 56–57.

Barros, T., Kuhlbrandt, W., 2009. Crystallisation, structure and function of plant light-harvesting complex II. Biochim. Biophys. Acta Biogeosci. 1787, 753–772.

Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L., 2012. The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure 20, 391–396. http://www.wwpdb.org/.

Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York (Chapter 12).

Cherezov, V., Rosenbaum, D.M., Hanson, M.A., Rasmussen, S.G., Thian, F.S., Kobilka, T.S., Choi, H.J., Kuhn, P., Weis, W.I., Kobilka, B.K., Stevens, R.C., 2007. High-resolution crystal structure of an engineered human β_2-adrenergic G protein-coupled receptor. Science 318, 1258–1265.

Cogdell, R.J., Isaacs, N.W., Freer, A.A., Howard, T.D., Gardiner, A.T., Prince, S.M., Papiz, M.Z., 2003. The structural basis of light-harvesting in purple bacteria. FEBS Lett. 555, 35–39.

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties. second ed. W. H. Freeman & Co., New York (Chapter 7).

Deisenhofer, J., Epp, O., Miki, K., Huber, R., Michel, H., et al., 1985. Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3 Å resolution. Nature 318, 618–624.

Filmore, D., 2004. It's a GPCR world. Modern Drug Discov. 7, 24–28.

Finkelstein, A.V., Ivankov, D.N., Dykhne, A.M., 2006. http://arxiv.org/abs/physics/0612139 (physics.bio-ph).

Frauenfelder, H., 2010. The Physics of Proteins. An Introduction to Biological Physics and Molecular Biophysics. Springer, New York (Chapter 13).

Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E., Downing, K.H., 1990. Model for the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. J. Mol. Biol. 213, 899–929.

Hille, B., 2001. Ion channels of excitable membranes. Potassium Channels and Chloride Channels. Sinauer, Sunderland, MA (Chapter 5), pp. 131–168.

Huber, R., 1989. Nobel Lecture. A structural basis of light energy and electron transfer in biology. EMBO J. 8, 2125–2147.

Jessell, T.M., Kandel, E.R., Schwartz, J.H., 2000. Principles of Neural Science, Ion Channels, fourth ed. McGraw-Hill, New York (Chapter 6), pp. 105–124.

King, N., Hittinger, C.T., Carroll, S.B., 2003. Evolution of key cell signaling and adhesion protein families predates animal origins. Science 301, 361–363.

Landau, L.D., Lifshitz, E.M., 1977. Quantum Mechanics. A Course of Theoretical Physics, vol. 3. Pergamon Press, New York (Chapter 3, Section 25).

Leonova, M.M., Fufina, T.Y., Vasilieva, L.G., Shuvalov, V.A., 2011. Structure-function investigations of bacterial photosynthetic reaction centers. Biochemistry (Mosc.) 76, 1465–1483.

Liu, Z., Yan, H., Wang, K., Kuang, T., Zhang, J., Gui, L., An, X., Chang, W., et al., 2004. Crystal structure of spinach major light-harvesting complex at 2.72 A resolution. Nature 428, 287–292.

Long, S.B., Tao, X., Campbell, E.B., MacKinnon, R., 2007. Atomic structure of a voltage-dependent K^+ channel in a lipid membrane-like environment. Nature 450, 376–382.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co., New York (Chapters 11, 12).

Neumann, S., Hartmann, H., Martin-Galiano, A.J., Fuchs, A., Frishman, D., 2012. CAMPS 2.0: exploring the sequence and structure space of prokaryotic, eukaryotic, and viral membrane proteins. Proteins 80, 839–857.

Overington, J.P., Al-Lazikani, B., Hopkins, A.L., 2006. How many drug targets are there. Nat. Rev. Drug Discov. 5, 993–996.

Pebay-Peyroula, E., Rummel, G., Rosenbusch, J.P., Landau, E.M., 1997. X-ray structure of bacteriorhodopsin at 2.5 Angstroms from microcrystals grown in lipidic cubic phases. Science 277, 1677–1681.

Rang, H.P., 2003. Pharmacology. Churchill Livingstone, Edinburgh, p. 60.

Roman, E.A., González Flecha, F.L., 2014. Kinetics and thermodynamics of membrane protein folding. Biomolecules 4, 354–373.

Skach, W.R., 2009. Cellular mechanisms of membrane protein folding. Nat. Struct. Mol. Biol. 16, 606–612.

Sprang, S.R., 1997. G protein mechanisms: insight from structural analysis. Annu. Rev. Biochem. 66, 639–678.

Stryer, L., 1995. Biochemistry, fourth ed. W.H. Freeman & Co., New York (Chapters 10-13, 15, 35-36).

Trzaskowski, B., Latek, D., Yuan, S., Ghoshdastider, U., Debinski, A., Filipek, S., 2012. Action of molecular switches in GPCRs—theoretical and experimental studies. Curr. Med. Chem. 19, 1090–1109.

Volkenstein, M.V., 1979. Electronic-conformational interactions in biopolymers. Pure Appl. Chem. 51, 801–829.

Lecture 13

Now we will focus on globular proteins or, rather, on water-soluble globular proteins.

They are the best-studied group: for many hundreds of them the spontaneous self-organization is known; for many thousands, their atomic 3D structure (with mutants and various functional states taken into account, these numbers are increased many-fold). Therefore, it is this type of proteins that is usually meant when "the typical protein structure," "the regularities observed in protein structure and folding," etc., are discussed (Volkenstein, 1977; Cantor and Schimmel, 1980; Creighton, 1993; Stryer, 1995; Branden and Tooze, 1991; Nelson and Cox, 2012).

After this remark, let us consider the structures of globular proteins (Figs. 13.1 and 13.2).

They have been yielded first by X-ray (Kendrew and Perutz, 1957; Perutz, 1992), and later by two- and many-dimensional NMR studies (Wüthrich, 1986) by hundreds of laboratories, and storied in the Protein Data bank (PDB) (Berman et al., 2012). PDB forms a basis for protein structure presentation, analysis, and classification (Murzin et al., 1995) used ubiquitously and, in particular, in this book. Many structures have been yielded by various firms for their own purposes.

> *Inner voice*: Does the structure seen by X-rays in the crystal coincide with the protein structure in solution?
>
> *Lecturer*: It virtually does, as a rule. This is supported by three groups of data. First, it can often be shown that a protein preserves its activity in the crystal form (see, eg, Vas et al., 1979; Fersht, 1999). Second, sometimes one protein can form different crystals, with its structure virtually unchanged. Finally, the NMR-resolved structure of a protein in solution and its X-ray-resolved structure in a crystal are virtually the same (Fig. 13.3). However, a reservation should be made that some flexible portions of proteins (some side chains, loops, as well as interdomain hinges in large proteins) may have a changed structure after or due to crystallization. But this only concerns either details of protein structures or connections between domains and subglobules in large proteins, rather than small single-domain protein globules, which usually are virtually solid.

I should add that X-rays see not only the "static," averaged structure of a protein (which is the subject of this lecture and the next) but also thermal vibrations of protein atoms, which will be discussed briefly later.

So, what is the bird's-eye view of water-soluble globular proteins?

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00013-2

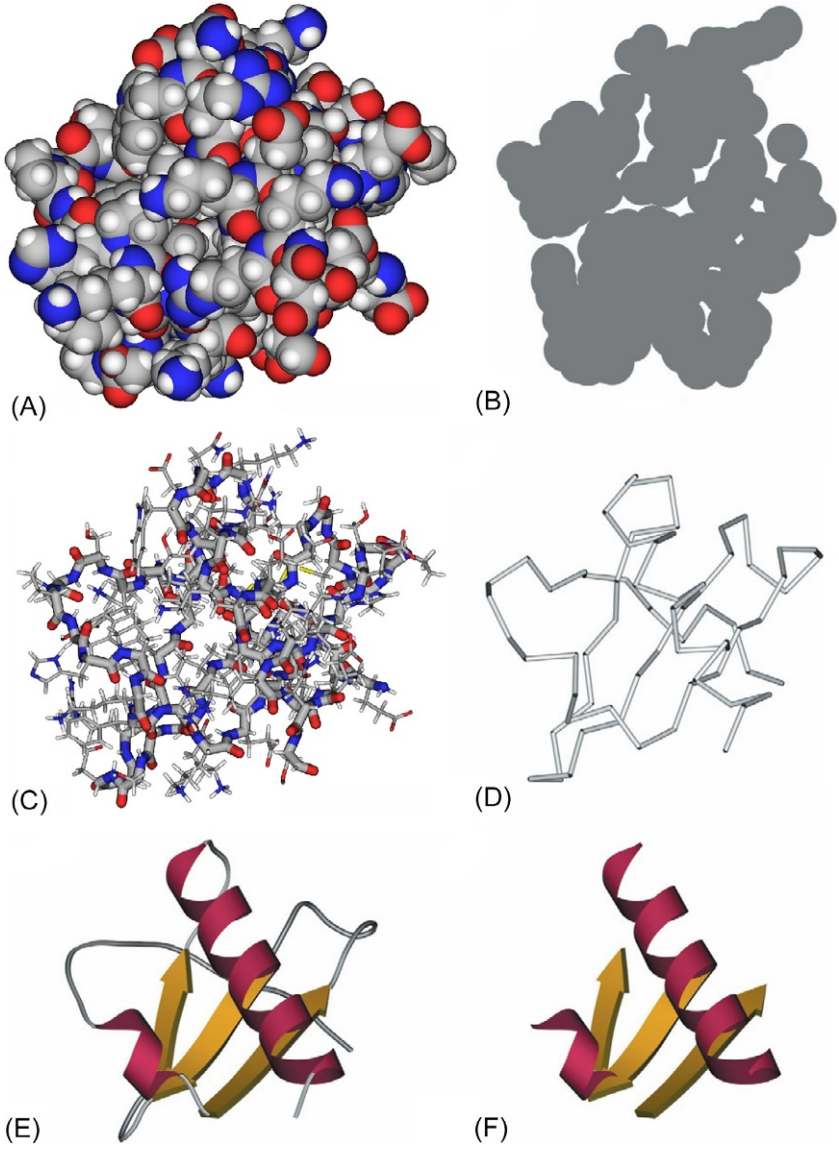


FIG. 13.1 The structure of a small protein, interleukin 8, shown in different ways. (A) The atomic model (nitrogens in blue, oxygens in red, carbons in gray, hydrogens in white); because of the close packing of the chain, we see only the protein surface. (B) The cross-section of the atomic model emphasizes the close packing. (C) Wire model of the main chain (dark line) and side chains (the lighter projections). (D) The pathway of the main chain. (E) Diagram of the protein fold showing the secondary structures involved (two α-helices and one β-sheet consisting of three β-strands). (F) Structural framework (stack) of the protein globule built up from secondary structures. The projection and scale are the same for all drawings.

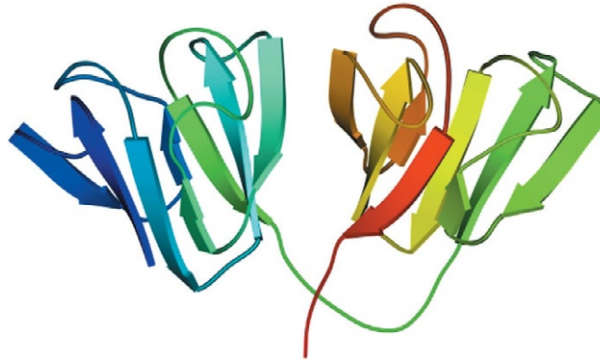


FIG. 13.2 Globular domains in γ-crystallin. The pathway of the chain is traced in the "rainbow" colors (from blue at the N-terminus via green in the middle to yellow and red at the C-terminus).

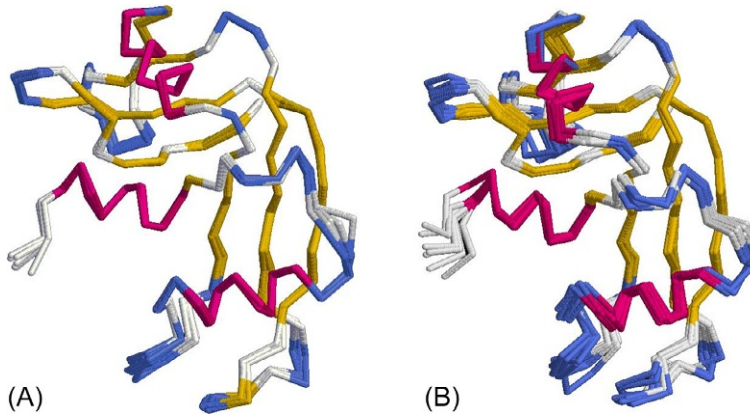


FIG. 13.3 Protein structure in crystal and in solution is nearly the same. (A) The best superposition of seven X-ray structures of the bovine ribonuclease A main chain in different crystals. All differences in these structures are small; the smallest ones are observed in α-helices (red) and β-structures (yellow); the largest in β-turns (blue) and irregular parts of the chains (white). (B) The same protein in solution: seven variants of the structure corresponding to one NMR experiment. The differences in these variants are somewhat larger than those in panel (A), which shows that the NMR solution structure is less accurate than the X-ray crystal structure.

We see that short chains (of 50–150 or, less frequently, 200–250 residues) pack into a compact 25–40 Å globule (Fig. 13.1), and that larger proteins consist of a few such subglobules, or domains (Fig. 13.2) (Schulz and Schirmer, 1979, 2013; Cantor and Schimmel, 1980; Creighton, 1993; Stryer, 1995; Branden and Tooze, 1991; Fersht, 1999; Nelson and Cox, 2012).

The protein chain is packed into a globule as tightly as organic molecules into a crystal. This is clear when you look at both the protein surface (Fig. 13.1A) and the cross-section of a protein globule shown in Fig. 13.1B. However, when examining a protein, we will not focus on the closely packed

electron clouds (or van der Waals surfaces) of atoms: then we cannot see the protein anatomy, that is, nothing can be seen inside the protein; instead, we will inspect the atom-"flesh"-free (Fig. 13.1C) and even side-chain-free (Fig. 13.1D) skeletons (wire models) of protein molecules. But do not submit to the impression (often created by drawings) that a protein molecule has a loose structure!

The frameworks of spatial structures of nearly all globules (domains) are composed of the regular secondary structures already familiar to us: α-helices and β-sheets (Fig. 13.1E) stabilized by regular H-bonds in the regular main chain. In globular proteins the total proportion of α- and β-structures amounts to 50–70% of the number of residues. By the way, Pauling, Corey, and Branson theoretically predicted these and β secondary structures prior to the resolution of atomic structures of protein molecules.

The arrangement of α-helices and β-sheets is not only the most fixed in each molecule (Fig. 13.3), and it is most nonvariable (more so than the position of irregular loops and tails chain) throughout the evolution (Fig. 13.4). Therefore, it is this arrangement determining the main features of protein architecture that forms a basis for structural classification of proteins.

The hydrophobic core (or cores) of the protein is surrounded by α- and β-structures, while irregular loops are moved towards the edge of the globule. The loops almost never enter the interior of the protein. This can be easily explained by the necessity for their peptide groups, uninvolved in secondary structures, to preserve their H-bonds to water, otherwise the globule's stability would be compromised. (Note that X-rays often find H-bonds between loops, α-helix ends and β-sheet edges and water molecules.)

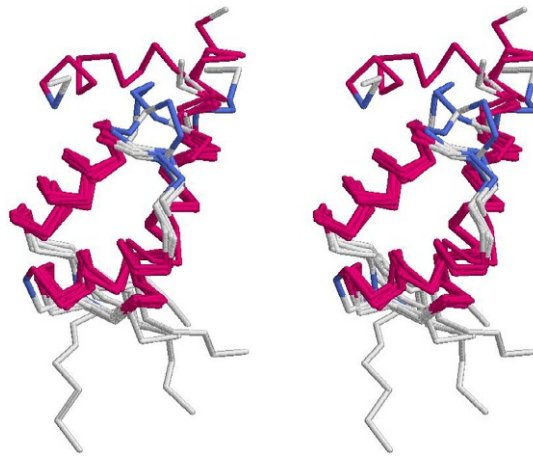


FIG. 13.4 Stereo drawing of the superposed structures of seven distantly related proteins (homeodomains of different eukaryotes, from yeast to mammals). The greatest similarity is observed in the general structural core mainly composed of α-helices (here, the difference is about 1 Å); in most of irregular regions the similarity is smaller, and at the ends of chains the differences are maximal.

The structural features of the main chain are the basis for subdivision of globular proteins into "pure" β-proteins, "pure" α-proteins, and "mixed" α/β and α+β-proteins. Strictly speaking, this classification refers to small proteins, as well as to separate domains (ie, to compact subglobules forming large proteins); large proteins can contain, say, both β- and α-domains.

Of particular interest to us are: (1) the architecture of packing of α- and β-structural segments into a compact globule (Fig. 13.1F) and (2) the pathway taken by the chain through the globule (Fig. 13.1E) or, as it is often called, "the topology of the protein globule" (taken together, the architecture and topology form a "folding pattern" of the protein molecule).

We will frequently use simplified models of protein structures (Fig. 13.5). The simplification implies not only focusing on secondary structures (with details of loop structures neglected) but also paying no attention to the difference in size of these structures or to details of their relative orientation (in this way, we pass from "folds" (Fig. 13.5A) to "folding patterns" (Fig. 13.5B) of protein chains).

The simplification is justified by the change in the details of loop structures and precise sizes and orientations of structural segments (and even of some small structural segments themselves) that occurs when the protein is compared with another one of similar sequence (ie, with its close relative of the same origin, Fig. 13.4)—eg, when hemoglobin α is compared with hemoglobin β (Fig. 13.6).

The next, higher level of simplification necessary for classification of protein structures is restricted to the packing of structural segments in a globule, that is, the stacks are composed of secondary structures, with no attention paid to loops connecting these secondary structures in a molecule (Fig. 13.5C).

I will purposely use such simplified models of chain folds and packing along with computer-produced "true" protein structures. It might seem

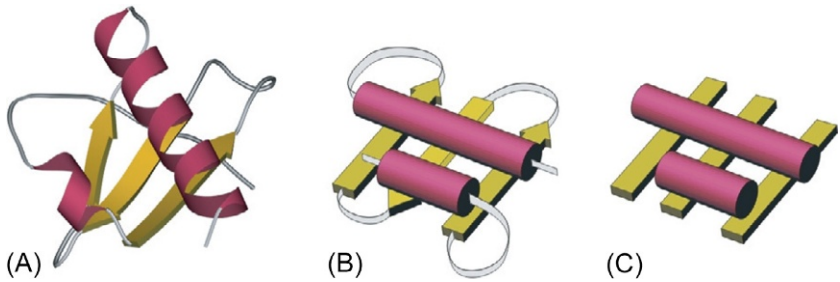


FIG. 13.5 Simplified models of protein structures. (A) A detailed *fold* describing the positions of secondary structures in the protein chain and in space (see also Fig. 13.1E). (B) The *folding pattern* of the protein chain with omitted details of loop pathways, the size and exact orientation of α-helices (shown as cylinders) and β-strands (shown as arrows). (C) *Packing*: a stack of structural segments with no loop shown and omitted details of the size, orientation and direction of α-helices and β-strands (which are therefore presented as ribbons rather than as arrows).

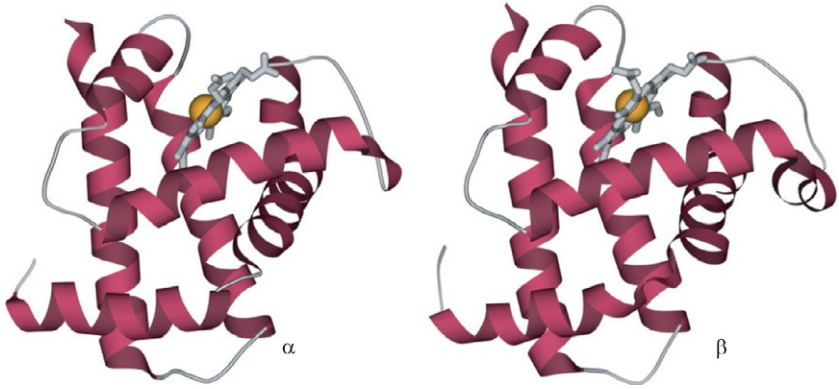


FIG. 13.6 Two close relatives: horse hemoglobin α and horse hemoglobin β (both possessing a heme shown as a wire model with iron in the center). Some differences are in details of loop conformations, in details of the orientation of some helices, and in one additional helical turn available in the β globin, on the right.

pointless to use the simplified models when a computer can describe the structure "as it is." However, this "as it is" picture has a lot of unnecessary details, while models embody the main features that are the same in similar proteins. Therefore, the models are useful both in classifying protein structures and in outlining their major typical features. When scrutinizing a picture of a protein, we cannot but outline its typical features in our minds—exactly what is done by models, which simply help viewers to systematize their intuitive perception. Besides, models allow us to compose the "verbal portrait" of a protein. Because these models and "verbal portraits" embody the main features and omit details, they will be of practical use as soon as you would like to find out how a protein under consideration resembles others. Of course the omitted details can be basic in protein functioning (as we will see later), but this only emphasizes that the function of a protein is relatively independent of the folding pattern of its chain.

Chain "packings" and "folding patterns" do not bring into focus all possible (loose, open-work, etc.) complexes of structural segments but only those closely packed. Thus, they outline the configuration areas corresponding to close (although free of steric overlapping) packing of the protein chain into a globule, that is, the vicinity of sufficiently deep energy minima of nonbonded interactions. These areas allow us not only to classify known protein structures but also to predict new ones yet to be detected (Levitt and Chothia, 1976; Richardson, 1977, 1981; Finkelstein and Ptitsyn, 1987; Chothia and Finkelstein, 1990).

It is not out of place to mention that, when speaking about classification of protein structures about similarities displayed, and so on, I will not mean the commonplace that all globins are alike irrespective of whether their host is a man or a lamprey—this is certainly true, and proteins can be divided into

phylogenetic classes within which their functions and, importantly, their amino acid sequences do not vary much. However, similar structural features are often intrinsic to proteins that, as revealed by numerous tests, evolutionarily have nothing in common. I will concentrate on this purely structural (not genetic) similarity.

We begin with the architecture of β-proteins. Structurally, β-structural domains turn out to be simpler than others: two (or sometimes several) β-sheets composed of extended chain segments are stacked one onto another. In other words, the "stacks" of secondary structures look quite simple in β-proteins. The *anti*parallel β-structure predominates in β-proteins.

Since proteins are composed of asymmetric (L) amino acids, the extended β-strands are slightly twisted individually: as you may remember, the energy minimum of an extended conformation is positioned above the diagonal in the Ramachandran plot (Chothia, 1973). The twisted β-strands are H-bonded into β-sheets that are arranged in a twisted propeller-like assembly (Fig. 13.7). The angle between adjacent extended strands of the β-sheet is about −25 degree. This propeller-like assembly looks *left*-handed when viewed across the β-strands (Fig. 13.7A) and *right*-handed when viewed along the β-strands (Fig. 13.7B). The latter is a common viewpoint (*along* the β-strands), and hence the β-sheet is said to have a *right*-handed twist.

There are two basic packing types for two β-sheets (Efimov, 1977; Chothia and Janin, 1981), namely, *orthogonal* packing and *aligned* packing (Fig. 13.8A and B). In both cases the sheets pack "face-to-ace" around a hydrophobic core of the domain, though their relative arrangement is different: in the second case the angle between the sheets is about −30 degree only (±10–15 degree), while in the first case it amounts to 90 degree (±10–15 degree); angles beyond these two ranges (specifically, angles of about +30 degree) are rare.

In *orthogonal* packing (Fig. 13.8A), the β-strands are twisted and usually slightly bent, so that the overall architecture of the "stack" resembles a cylinder

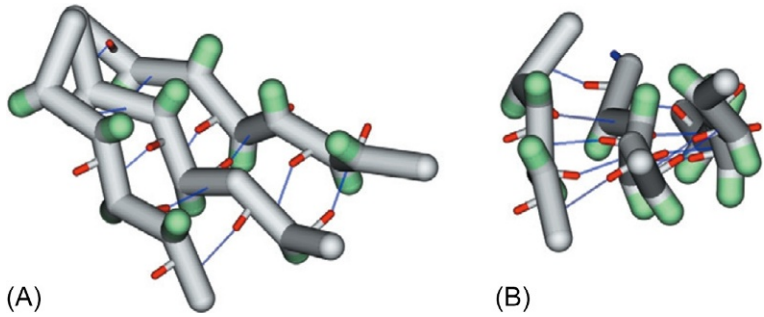


FIG. 13.7 The β-sheet as viewed (A) across and (B) along its β-strands. The sheet is pleated (which is emphasized by projecting the C^{β}-atoms shown in green) and usually, like here and in the majority of most β-proteins shown below, has a right-handed (viewed along the strands) propeller-like twist. H-bonds between β-strands are shown by thin lines.

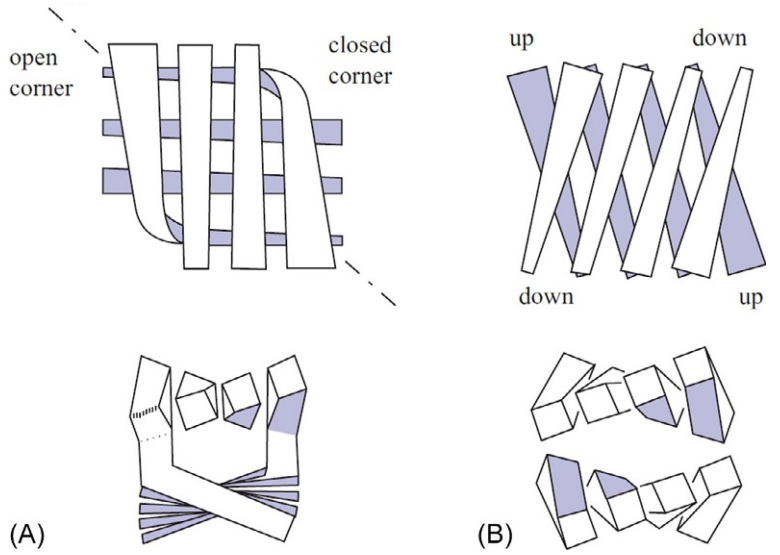


FIG. 13.8 The orthogonal (A) and aligned (B) packing of β-sheets viewed face on (top) and from their lower end (bottom). In the face view, the β-strands are wider as they approach the viewer. The dashed line shows the axis of the orthogonal β-barrel to which both "open" corners belong. Here the two β-sheets are most splayed. At the two "closed" corners the sheets are extremely close together; here the chain bends and passes from one layer to the next. In the orthogonal packing the hydrophobic core is almost cylindrical. In contrast, in the aligned packing, the core is flat, the distance between the twisted sheets remains virtually unchanged, and the relative arrangement of the twisted sheets allows the hydrophobic faces of twisted β-strands to make contact over a great length. *(Adapted from Chothia, C., Finkelstein, A.V., 1990. The classification and origins of protein folding patterns. Ann. Rev. Biochem. 59, 1007–1039.)*

with a significant angle between its axis and the β-strands. This type of β-sheet packing is often called the β-cylinder or the β-barrel, although in β-cylinders composed of *anti*parallel β-structures (unlike in those of parallel β-structures to be discussed later) the two β-sheets are usually clearly distinguished, because on opposite sides of the barrel the H-bond net is often fully or partially broken. In the "closed" corner of this packing, β-segments of both sheets come close together, which enables the chain to pass from one sheet to the other at the expense of a 90 degree bend; it can be said that a single sheet is bent with one part imposed on the other. At the opposite (open) corners, the β-sheets splay apart, the open corner being filled with either an α-helix or irregular loops (Chothia and Janin, 1981), or even with an active site, as in the case of retinol-binding protein (Fig. 13.9).

The *aligned* packing (Fig. 13.8B) is typical of nonbent sheets with a propeller-like twist. This packing type is usually called a β-sandwich. Its ends are covered with irregular loops protruding from the ends of the β-strands (as seen in Fig. 13.10). In some β-sandwiches the edge strands of the β-sheets

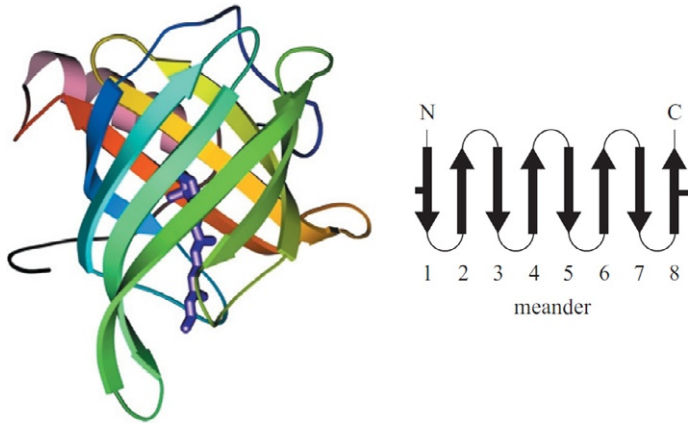


FIG. 13.9 Retinol-binding protein exemplifies β-sheet orthogonal packing. The chain pathway resembles the "meander" pattern (Richardson, 1977) (see the topological diagram, ie, the planar presentation of the β-sheet, on the right). In this diagram β-strands are shown as arrows. The "meander" results from the fact that β-strands adjacent in the chain are also adjacent at the surface of the cylinder; there are H-bonds between them (the H-bonding between the edge (in the planar diagram) β-strands is shown by small thick lines). The retinol-binding site is at the cylinder axis. Retinol is shown in violet. The numbers in the topological diagram reflect the order of structural segments in the chain.

are so close (and sometimes even H-bonded) that the packing acquires the shape of a cylinder with a small angle between its axis and the β-strands.

It should be stressed that the folding patterns are much less numerous than the known protein globules, and the types of stacks are in turn much fewer than the folding patterns of the protein chain. One and the same stack, that is, one and the same packing of structural segments, can conform to various patterns of chain folding in a globule; in other words, these segments can be arranged in the polypeptide chain in different ways.

Fig. 13.10 shows how one and the same β-sandwich (with a variety of topologies, that is, with different pathways taken by the chain through this sandwich) serves as the structural basis of three distinct protein domains: (a) a domain of γ-crystallin, (b) the β-domain of catabolite activating protein, and (c) the virus coat protein. The folding patterns of the last two proteins are the same, that is, in these proteins the chain takes the same pathway through identical β-structural stacks (this is emphasized by the brace in Fig. 13.10).

In continuation of this line, Fig. 13.11 shows how one and the same β-cylinder (with two different chain topologies, ie, with two different pathways taken by the chain through the orthogonal packing of the β-sheets) serves as the basis for both a serine protease like chymotrypsin (a) and an acid protease like pepsin (b).

Examples of such structural similarity in the absence of any other apparent relationship between the proteins are plentiful.

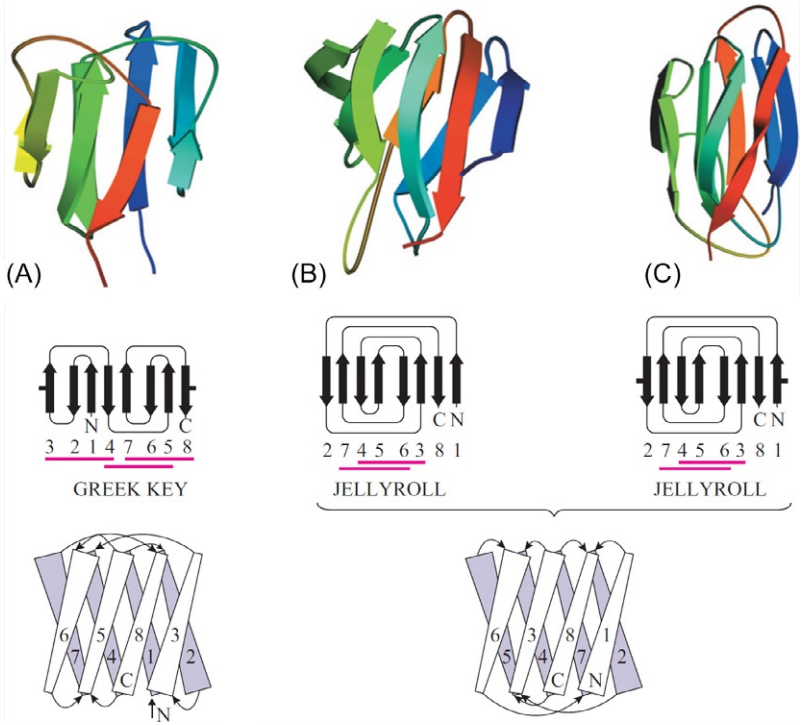


FIG. 13.10 Examples of β-sheet aligned packings. The chain folding patterns for (A) γ-crystallin (see also Fig. 13.2), (B) the β-domain of catabolite activating protein and (C) the coat protein of satellite tobacco necrosis virus. The topological diagrams for these proteins are shown below. For all these proteins, the topology contains a "Greek key" (Richardson, 1977) (a "long hairpin bent in two") where four β-strands are adjacent in the chain and antiparallel, and there are H-bonds between the first and the fourth strand. In the topological diagrams, structures of the two-layer β-sandwich are drawn in one plane (as for a cylinder cut along its side and unrolled flat). The place of cutting is chosen to stress the symmetry of the chain fold. For example, in γ-crystallin, the slit between β-strands 3 and 8 stresses the similarity of the first (strands 1–4) and the last (strands 5–8) halves of the domain. A shorter distance between the β-strands implies their H-bonding; H-bonds (if any) between the edge β-strands are shown as small thick projecting lines. A gap between the β-strands separates two β-sheets of the sandwich (if there are no H-bonds between them). The domain of γ-crystallin contains a repeated Greek key: one formed by strands 1–4, and the other by strands 5–8; still another Greek key is composed of the β-strands 4–7. The proteins shown in drawings (B) and (C) have Greek keys formed by strands 3–6 and 4–7. Moreover, their topology can be described as a repeatedly bent hairpin (usually called a "jellyroll") where β-strand 1 is H-bonded to 8, and 2–7, in addition to the H-bonds between strands 3 and 6, 4 and 5, typical of Greek keys. Note that for the purpose of enveloping the globule core by the chain a Greek key proves to be better than "meander" topology (Fig. 13.9) because apart from the β-structure on the sides, it provides enveloping by loops from below and above. Usually, β-sheet aligned packings are β-sandwiches (A,B), but some of them (eg, the coat protein of satellite tobacco necrosis virus and coat proteins of some other viruses) can also be seen as β-cylinders with colinear β-strands (C).

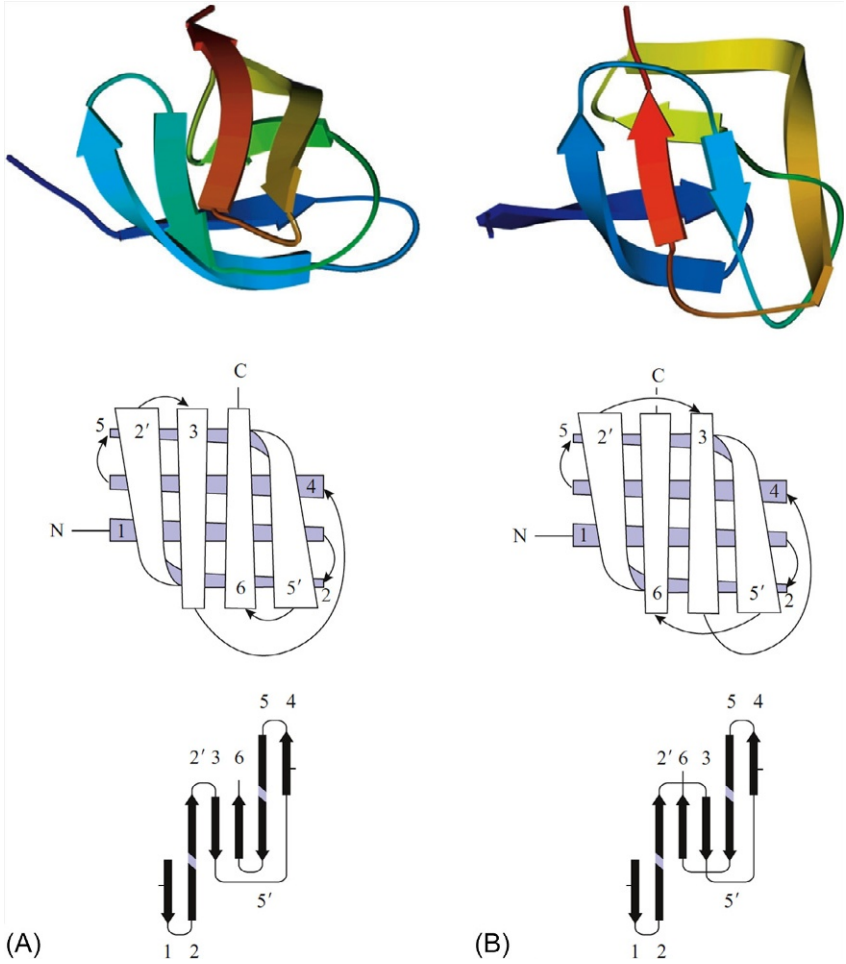


FIG. 13.11 The chain folding patterns in a serine protease such as chymotrypsin (A) and in an acid protease such as pepsin (B). For the latter, the loops are shortened and rather schematic. The orthogonal packings of β-sheets in these proteins are shown along with the β-sheet topology diagrams. In both folding patterns the β-sheets has a bend, such that their edges move away from the reader and stick together by H-bonding (marked as short lines); the places where the β-sheet bends are colored lighter in the topology diagrams (see β-strands 2 and 5).

Here I cannot resist the temptation of showing you another β-sandwich-based protein. It is immunoglobulin; each domain of this large protein is arranged in (or close to) the way shown in Fig. 13.12.

Folding patterns of this type are intrinsic to the chains of about 50 other superfamilies, bearing no sequence similarity to immunoglobulin (although some also are responsible for specific binding to certain agents, eg, in the course of cell recognition).

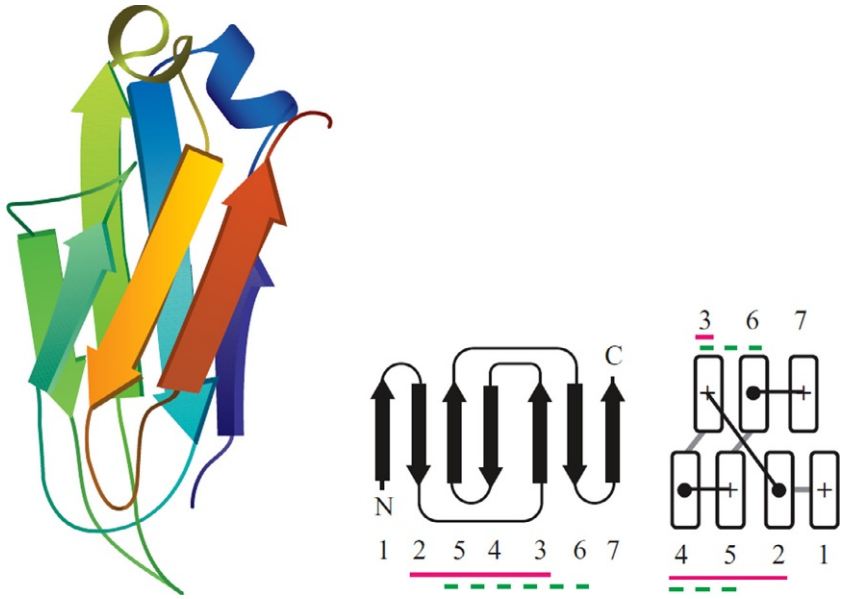


FIG. 13.12 The aligned packing of β-sheets in the constant domain of the light chain of immunoglobulin κ. On the left, a detailed diagram of the protein is shown; the chain pathway is traced in color (rainbow from blue to red) from the N- to the C-terminus. The topological diagram (in the center) accentuates the "Greek keys." On the right, the protein is shown as viewed from below (ie, from the butt-ends of structural segments; the butt-ends shown as rectangles). The cross corresponds to the strand's N-end (ie, "the chain runs from the viewer"), and the dot to the C-end (ie, "the chain runs towards the viewer"). The segment-connecting loops close to the viewer are shown by black lines, and those distant (on the opposite side of the fold) by light lines. Note that such a diagram allows the presentation of the colinear packing of these segments (β-strands) in the simplest possible way. It also visualizes the spatial arrangements of "Greek keys" and makes evident that two of them (formed by strands 2–5 and 3–6, respectively) differ in their spatial arrangements. The structure formed by two Greek keys that overlap as shown in this figure is sometimes called a "complete Greek key."

Some members of these superfamilies are somewhat different from the "standard" structure presented in Fig. 13.12. So, in some proteins β-strand 1 forms a parallel β-structure with strand 7 (and then it can even lose contact with strand 2); in others β-strand 4 moves to strand 3 from strand 5; sometimes an additional β-hairpin forms in the connection between β-strands 3 and 4. But the core of the fold, which involves β-strands 2, 3, 5, 6, 7, remains unchanged.

Apart from the pleasure of showing you this very popular folding pattern, I also aim to show that the easiest way to illustrate the folding pattern of a protein with more or less colinear packing of its structural segments is to use the diagram giving a view from the butt-ends of β-strands.

So far, we have considered the most significant "basic" arrangements of β-proteins. However, there are other "basic arrangements," for example, the "multiple-blade propeller."

In the neuraminidase "propeller" (Fig. 13.13), six inclined β-sheets form a rosette (in other proteins of this kind there may be as many as eight sheets).

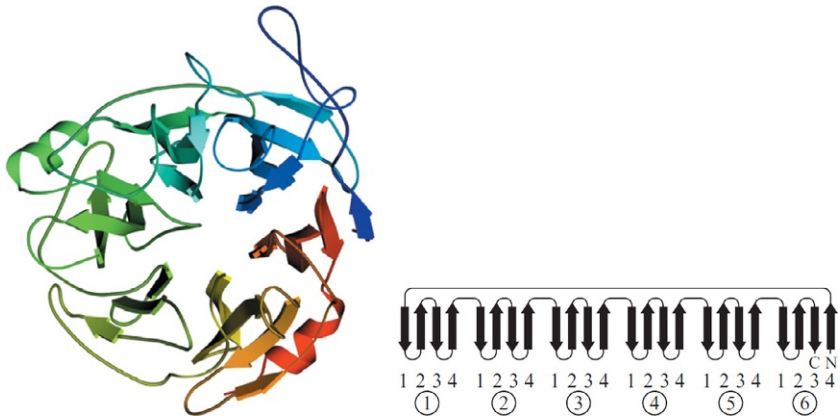


FIG. 13.13 The β-structure in the form of a "six-blade propeller" in neuraminidase, and a topological diagram of this protein, which is composed of six antiparallel β-sheets. In some other proteins, "blades" are formed by separate chains. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York, London (Chapter 5), with minor modifications.)*

If considered in pairs, the sheets form β-sandwiches; therefore the "propeller" can be described as a supercylinder built up from β-sandwiches.

As you can see, the axis is not covered with loops and the "indent" at the supercylinder axis contains the active site. We have already seen one similar position of the active site: in the retinol-binding protein (Fig. 13.9), it is also located in an indent in the middle of the cylinder; and we will see it again later.

The structural arrangement of the "β-prism" (also called the "β-helix") type (Fig. 13.14) is of interest mostly due to its regularity. The three facets of this prism are formed by three β-sheets (note: parallel β-sheets) such that the chain

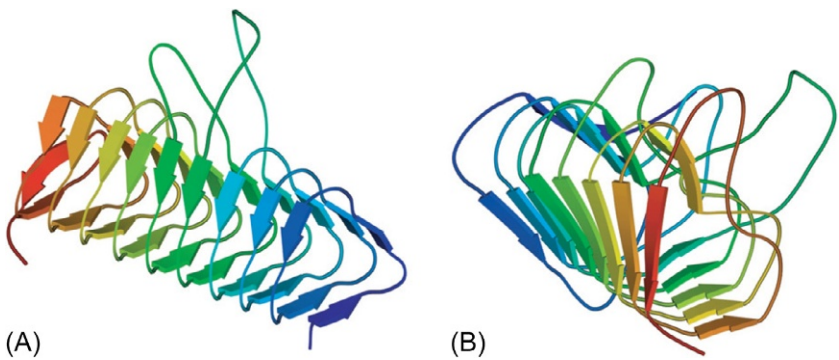


FIG. 13.14 The β-prism in acyl transferase (A) and in pectate lyase (B). Notice the handedness of the chain's coiling around the axis of the prism: it is unusual, *left*, in (A) and common, *right*, in (B). Also note that when the chain's coiling is uncommon, that is, left-handed as in (A), the common twist of the β-sheet is absent. This common twist, that is, the right-handed (viewed along the β-strands) propeller twist, is seen also in Figs. 13.7–13.13.

takes its pathway through them continuously passing from one sheet to the next. The chain appears to coil around the axis of this prism and forms either a right-handed helix, which is typical for joining parallel β-strands, or (in the other prism) a left-handed helix, which is extremely rare for the case of joining β-strands in other proteins. It is worth noting that in some primitive fishes (eg, lampreys) the immune role is performed by proteins built not on a classical immunoglobulin fold (Fig. 13.12) but on β-prisms (Fig. 13.14).

It would not be out of place here to discuss the *topology* of β-proteins. We have observed that β-proteins are built up from mostly antiparallel β-structures. The structure of the majority of β-proteins that have been discussed so far is purely antiparallel. Although sometimes a minor admixture of the parallel structure was observed (see Fig. 13.11B), proteins built up from a purely parallel β-structure are extremely rare, although they do exist (see Fig. 13.14).

The fact that an admixture of parallel and antiparallel structures rarely occurs is not surprising since parallel and antiparallel β-structures have somewhat different conformations, and therefore their connection is likely to be energetically unfavorable. The extent of correlation between the unfavorability and the uncommon occurrence of various structures will be discussed in a later lecture, but in principle, it is clear that a stable system (such as protein) must be composed of mostly stable elements and avoid those internally unstable.

The mostly *anti*parallel character of β-sheets in β-proteins is closely connected with the fact that their architecture is usually based on β-hairpins (Fig. 13.15). These hairpins are often bent and sometimes may even have two or three such bends (see Fig. 13.15 and also Fig. 13.10B and C).

The pathways of loops connecting β-segments usually start and finish on the same edge of the fold (ie, the loops do not cross the "stack" but cover its butt-end). This is well seen from almost all drawings. The loops, even long ones, tend to connect ends of β-segments that are close in space. That is why, as a rule, β-segments adjacent in the chain are not parallel and tend to form antiparallel β-hairpins.

Also note that "overlapping" loops (or "crossed loops") occur rarely (an exception of this kind is shown in Fig. 13.11B), probably because one of the crossed loops must have an energetically unfavorable additional bend (to avoid a collision or dehydration). The avoidance of loop overlapping is a general structural rule for proteins (Lim et al., 1978; Ptitsyn and Finkelstein, 1979).

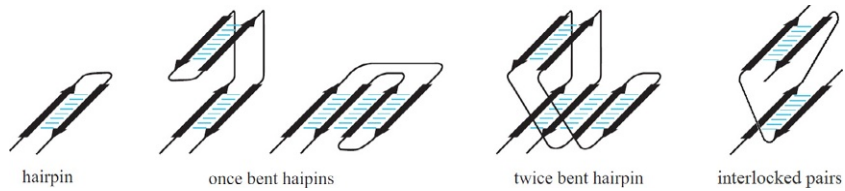


FIG. 13.15 Antiparallel β-hairpins (bent β-hairpins are typical of edges of β-sandwiches) and interlocked pairs of β-strands typical of middles of β-sandwiches (Kister et al., 2002).

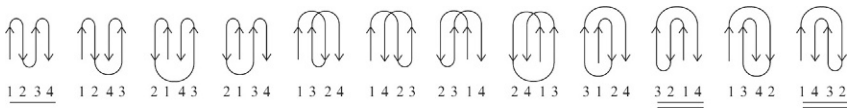


FIG. 13.16 Possible topologies of sheets composed of four β-strands. The scheme includes only the sheets where two β-strands adjacent in the chain are oppositely directed. Among these, the common topologies are "meander" (underlined once) and two "Greek keys" (underlined twice), the latter two being different only in the direction of the chain turn from the hairpin consisting of strands 1 and 4 to the hairpin consisting of strands 3 and 2. The "meander"-containing protein is exemplified by retinol-binding protein (see Fig. 13.9); the examples of "Greek key"-containing proteins are γ-crystallin and other proteins shown in Fig. 13.10, or trypsin (Fig. 13.11).

Among a variety of configurations of a β-sheet formed by a continuous chain (Fig. 13.16), those really abundant are only three: two "Greek keys" and the "meander" (underlined in Fig. 13.16)—by the way, "Meander" is the name of a very winding river in Ancient Greece (now Turkey)—in the meander pattern, β-strands adjacent in the chain are also adjacent in space (Figs. 13.10 and 13.16) and usually linked with H-bonds. The "popular" folding patterns have none of the disadvantages mentioned above: a mixture of antiparallel and parallel β-structures and the crossed loops.

It is characteristic of the "Greek key" pattern (which can be seen on ancient vases and garden railings), that four β-strands adjacent in the chain are antiparallel, and that there are H-bonds between the first and the fourth strands. Actually, the second and/or the third β-strand of the "Greek key" usually belong to another β-sheet rather than to the same one (as it may appear from Fig. 13.16). This gives rise to various spatial structures (the so-called *Efimov's* "*abcd*" *structures*) with the same Greek key topology but with different shapes in space (Efimov, 1995) (Fig. 13.17). Look for them in Figs. 13.10–13.12.

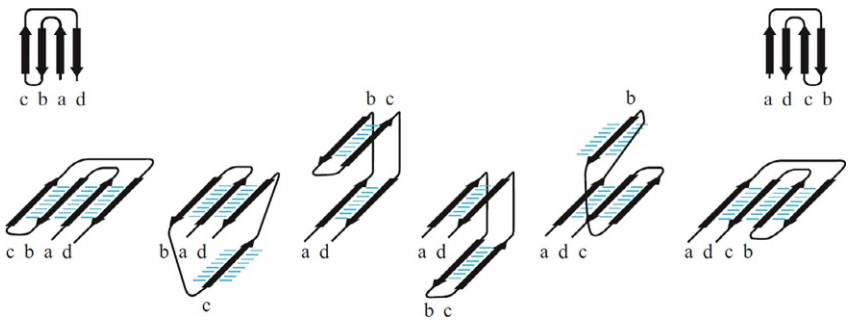


FIG. 13.17 Several kinds of the supersecondary structure: various spatial Efimov's "*abcd*" structures with the Greek key topology. Notice the *right-handedness of the superhelix* that consists of two parallel β-strands from one β-sheet and one β-strand (between them) from another β-sheet. It is seen, for example, in the superhelices b-c-d (the second drawing from the left in the lower line) and a-b-c (the second drawing from the right in the lower line). The "right-handed" connection of parallel strands of the same β-sheet is typical for proteins (Nagano, 1973; Richardson, 1976); the reverse (left-handed) is extremely rare.

These typical protein structures (hairpins, meanders, Greek keys, *abcd* structures, etc.), which are built up from elements of the β- (and/or α-) structure that are adjacent in the chain, are often called "supersecondary" structures.

REFERENCES

Berman, H.M., Kleywegt, G.J., Nakamura, H., Markley, J.L., 2012. The Protein Data Bank at 40: reflecting on the past to prepare for the future. Structure 20, 391–396. http://www.wwpdb.org/.

Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Publishing Inc., New York, London (Chapter 5).

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman & Co., New York (Part 1, chapters 2, 9; part 2, chapter 13).

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

Chothia, C., Finkelstein, A.V., 1990. The classification and origins of protein folding patterns. Annu. Rev. Plant Physiol. Plant. Mol. Biol. 59, 1007–1039.

Chothia, C., Janin, J., 1981. Relative orientation of close-packed beta-pleated sheets in proteins. Proc Natl Acad Sci U S A 78, 4146–4150.

Creighton, T.E., 1993. Proteins: Structures and Molecular Properties, second ed. W. H. Freeman & Co., New York (Chapter 6).

Efimov, A.V., 1977. Stereochemistry of the packing of alpha-spirals and beta-structure into a compact globule. Dokl. Akad. Nauk. SSSR (in Russian) 235, 699–702.

Efimov, A.V., 1995. Structural similarity between two-layer α/β and β-proteins. J. Mol. Biol. 245, 402–415.

Fersht, A., 1999. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. W. H. Freeman & Co., New York (Chapter 1).

Finkelstein, A.V., Ptitsyn, O.B., 1987. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50, 171–190.

Kendrew, J.C., Perutz, M.F., 1957. X-ray studies of compounds of biological interest. Annu. Rev. Biochem. 26, 327–372.

Kister, A.E., Finkelstein, A.V., Gelfand, I.M., 2002. Common features in structures of sandwich-like proteins. Proc. Natl. Acad. Sci. U. S. A. 99, 14137–14141.

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

Lim, V.I., Mazanov, A.L., Efimov, A.V., 1978. Stereochemical theory of the 3-dimensional structure of globular proteins. I. Highly helical intermediate structures. Mol. Biol. (in Russian) 12, 206–213.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.

Nagano, K., 1973. Logical analysis of mechanism of protein folding. I. Prediction of helices, loops and β-structures from primary structure. J. Mol. Biol. 75, 401–420.

Nelson, D.L., Cox, M.M., 2012. Lehninger Principles of Biochemistry, sixth ed. W.H. Freeman & Co., New York (Chapters 3–6, 27).

Perutz, M.F., 1992. Protein Structure. New Approaches to Disease and Therapy. W.H. Freeman & Co., New York.

Ptitsyn, O.B., Finkelstein, A.V., 1979. Prediction of Protein Three-Dimensional Structure by Its Amino-Acid Sequence. In: 12th FEBS Meeting (Dresden, 1978). 52. Pergamon Press, New York, pp. 105–111.

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

Richardson, J.S., 1977. β-Sheet topology and the relatedness of proteins. Nature 268, 495–500.
Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34, 167–339.
Schulz, G.E., Schirmer, R.H., 1979, 2013. Principles of Protein Structure. Springer, New York (Chapter 5).
Stryer, L., 1995. Biochemistry, fourth ed. W.H. Freeman & Co., New York (Chapter 2).
Vas, M., Berni, R., Mozzarelli, A., Tegoni, M., Rossi, G.L., 1979. Kinetic studies of crystalline enzymes by single crystal microspectrophotometry. Analysis of a single catalytic turnover in a D-glyceraldehyde-3-phosphate dehydrogenase crystal. J. Biol. Chem. 254, 8480–8486.
Volkenstein, M.V., 1977. Molecular Biophysics. Academic Press, London (Chapter 4).
Wüthrich, K., 1986. NMR of Proteins and Nucleic Acids. Wiley-Interscience, New York.

This page intentionally left blank

Lecture 14

Now we pass to α-proteins, which are proteins built up from α-helices. They are more difficult to classify than β-proteins (Levitt and Chothia, 1976). The reason is that the arrangement of β-strands in the sheets is stabilized by hydrogen bonds of the main chain (which is the same everywhere), while the arrangement of α-helices in a globule is maintained by close packing of their side chains, which vary greatly in size. This is why, unlike β-strands, α-helices do not pack into more or less standard sheets.

The α-proteins composed of long α-helices have the simplest structure. This structure is a bundle formed by almost parallel or antiparallel (in a word, co-linear) long α-helices. We have already met such bundles when considering fibrous and membrane proteins.

Fig. 14.1 presents three 4-helix proteins. These proteins, structurally very close, have different functions: cytochrome binds an electron, hemerythrin binds oxygen, and tobacco mosaic virus coat protein binds molecules that are much greater in size—other coat proteins and RNAs (Branden and Tooze, 1991). The first two proteins may have something in common in their function because they both act as carriers in the respiratory chain. This resemblance is far from close, though: in cytochrome, the polypeptide binds the heme that binds an iron that binds an electron, while in hemerythrin the polypeptide binds iron ions without any mediating heme, and two iron ions bind an oxygen. Thus, functionally, hemerythrin and cytochrome bear some resemblance, although only a minor one, and they have no common function with the RNA-binding virus coat protein—in spite of the fact that all three of them are very similar structurally. The similarity is not restricted to the overall architecture (a 4-helix bundle) but also involves the pathway taken by the chain through this bundle, that is, the folding pattern of the chain. The latter is well illustrated by a common topological diagram (Fig. 14.1, inset) that shows a view along the bundle axis.

Thus, proteins with the same folding pattern may have utterly different functions; the same we have seen for β-proteins. In contrast, hemerythrin and the classical oxygen-binding protein myoglobin (Fig. 14.2) have identical functions (the former in worms and the latter in vertebrates, including ourselves), while their architectures are utterly different except that they are both α-proteins. However, in hemerythrin all the α-helices are parallel, while in myoglobin they are assembled into two perpendicular layers. This is another example showing that proteins with different architectures can carry out similar functions, while similarly arranged proteins may have different duties.

Again, I am drawing your attention to nontrivial cases of the lack of relationship between protein structure and function, because undoubtedly you know

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00014-4

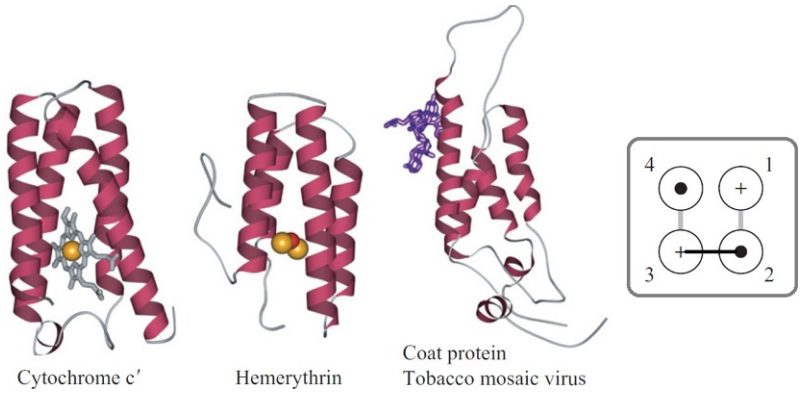


FIG. 14.1 Three α-proteins that are similar in architecture (4-helix bundle) but different in function: cytochrome *c′*, hemerythrin, and tobacco mosaic virus coat protein. Both the protein chain and co-factors are shown: wire models represent the heme (in cytochrome) and an RNA fragment (in virus coat protein), orange balls are for iron ions (in the cytochrome heme and in hemerythrin), and the red ball is for iron-bound oxygen (in hemerythrin). The overall architecture of such "bundles" resembles the co-linear packing of β-sheets. The topological diagram (inset) shows all these proteins as viewed (in the same orientation) from their lower butt-ends. The circles represent the ends of α-helices. The cross corresponds to the N-end of the segment (ie, the segment goes away from the viewer); the dot corresponds to its C-end (ie, the segment comes towards the viewer). The loops connecting the structural segments are shown by the black line (if the loop is close to the viewer) and by the light line (if it is on the opposite side of the fold). The numerals indicate the order of structural elements in the chain (from the N- to the C-terminus).

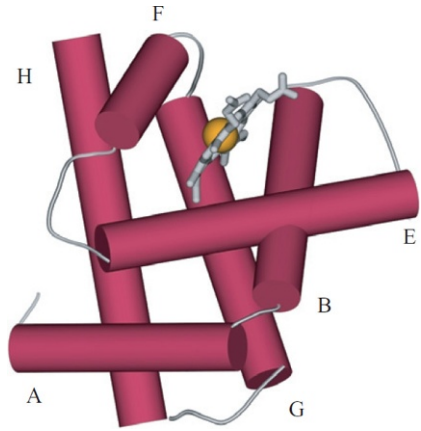


FIG. 14.2 The structure of globin: crossed layers of three α-helices each. The helices A, E, and F (lettered in accordance with their sequence positions) belong to the upper layer, while H, G, and B to the lower layer. The short helices (of 1–2 turns each) C and D are not shown since they are not conserved in globins. A crevice in the upper layer houses the heme. Such "crossed layers" resemble the orthogonal packing of β-sheets (the orthogonal contact of helices B and E is especially close, since both helices have glycine-formed dents at the contact point).

that the kindred proteins (eg, myoglobin and other globins) are similar in both architecture and function.

Besides, by comparing myoglobin with hemerythrin, I want to draw your attention to the fact that in both cases the active site (for the former, it is the heme with an iron ion inside; for the latter, two iron ions) is localized in the "architectural defect" of the protein structure, namely, in a crevice between the helices.

The "bundles" described above are typical of quite long α-helices. They are observed in water-soluble globular proteins, as well as in fibrous and membrane proteins. The protein core enveloped by the helices has an elongated, quasi-cylindrical shape. It is hydrophobic in water-soluble globular and fibrous proteins and hydrophilic in membrane ones.

The "crossed layers" (Fig. 14.2) are also formed by rather long helices; they have a flat hydrophobic core.

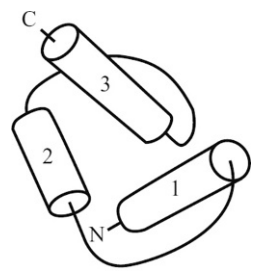
However, relatively short helices are more typical of globular proteins. For these helices (with a length of about 20 Å), a quasi-spherical packing around a ball-like core is more typical (Murzin and Finkelstein, 1983, 1988).

Fig. 14.3 illustrates a typical packing of helices in a globular protein. This packing cannot be described in terms of a parallel or perpendicular helical arrangement because the angles between helices are usually 40–60 degrees.

But even such intricate packings can be described and classified accurately enough using the "quasi-spherical polyhedron model" (Murzin and Finkelstein, 1983, 1988). For example, let us see (Fig. 14.4) how this model describes the α-helical globule of Fig. 14.3. Incidentally, these two figures are from a review of this work by A.G.M. and A.V.F., published in *Nature* by Chothia (1989).

I will not deny myself the pleasure of showing you another pair of figures from the same review (Fig. 14.5A).

By the way, the quasi-spherical polyhedron model is also sufficiently good for describing rather long helical bundles (such as those we saw in Fig. 14.1). This is illustrated by Fig. 14.5B.



Actinidin

FIG. 14.3 Typical packing of helices in a globular protein exemplified by the N-terminal domain of actinidin (the loops are traced very schematically). Note that the architecture of this domain cannot be described in terms of colinear and orthogonal packings of α-helices. *(Adapted from Chothia, C., 1989. Polyhedra for helical proteins. Nature 337, 204–205.)*

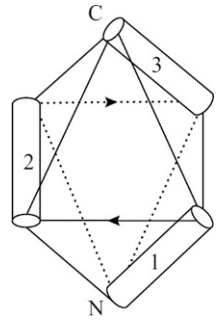


FIG. 14.4 The α-helix positions on the ribs of a quasi-spherical polyhedron that models the N-terminal domain of actinidin shown in Fig. 14.3. The shortcuts for helix-connecting loops are shown by arrows. *(Adapted from Chothia, C., 1989. Polyhedra for helical proteins. Nature 337, 204–205.)*

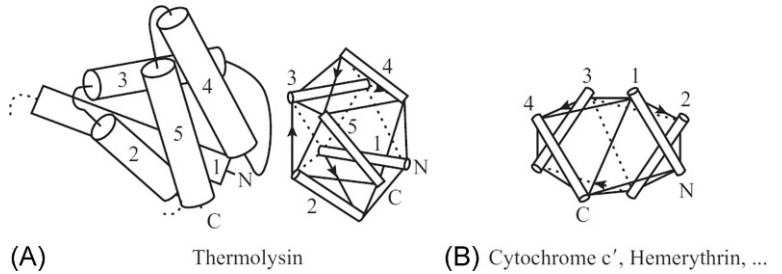


FIG. 14.5 More examples showing how the geometry of helix packings in globular proteins can be described by the quasi-spherical polyhedron model. (A) The C-terminal domain of thermolysin and its model showing the helix positions on the polyhedron ribs. (B) The model for the four-helix globule presented in Fig. 14.1. *((A) Adapted from Chothia, C., 1989. Polyhedra for helical proteins. Nature 337, 204–205.)*

Essentially, the quasi-spherical polyhedron model focuses on the positioning of α-helices packed around the ball-like core of the globule (Fig. 14.6).

The model only takes into account that α-helices, solid extended particles, surround the core closely, and that the polar helix ends are located on the globule surface. The geometry of any helix packing can be described by a polyhedron (Fig. 14.6) where each vertex corresponds to half of the helix. The most compact "quasi-spherical" polyhedra (Fig. 14.7) describe compact globules. The helix packings actually observed in globular α-proteins are close to these ideal packings. For a given number of helices, there is only one most compact polyhedron; it allows for a number (from 2 to 10) variants of helix positioning on the ribs of this polyhedron. The previously considered "helix bundles" "crossed layers" are among these arrangements.

Interestingly, in the observed architectures of α-proteins, it is not only helices that fit on the ribs of quasi-spherical polyhedra but also, as a rule, the helix-connecting irregular loops (see Figs. 14.4 and 14.5). In other words, a typical

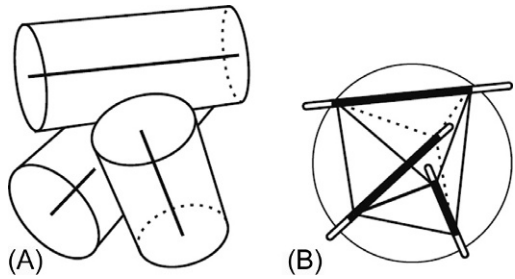


FIG. 14.6 This figure illustrates how the geometry of helix packings can be described by a polyhedron. (A) Three packed helices are shown as cylinders of diameter 10 Å (their axes are also shown). (B) To construct the polyhedron, a sphere of radius 10 Å is drawn from the center of the packing; the polyhedron vertices occur at its intersection with the helix axes. The sections of the ribs enclosed by the sphere are shown as dark lines. Each vertex corresponds to one-half of one helix. The helix axes form one set of the ribs of the polyhedron; it is completed by another set of ribs formed by connections linking the helix ends. *(Adapted from Murzin, A.G., Finkelstein, A.V., 1988. General architecture of α-helical globule. J. Mol. Biol. 204, 749–770.)*

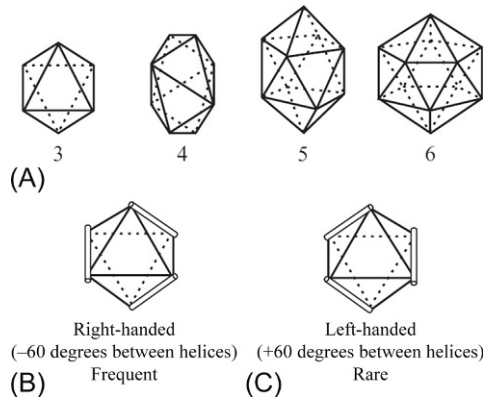


FIG. 14.7 (A) Quasi-spherical polyhedra describe the compact packing of three, four, five and six helices. Larger assemblies of helices cannot be placed around a single spherical core without screening the polar ends from water. Each polyhedron simultaneously describes several packing arrangements, ie, several types of "stacks" of helices; the stacks differ in helix positioning on the polyhedron ribs. For example, three helices form two different arrangements: (B) a right-handed bundle (as that in Figs. 14.3 and 14.4); (C) a left-handed bundle. Four helices form 10 arrangements, five helices form 10 arrangements, and six helices form eight arrangements ["stacks" for four-, five- and six-helix globules are not shown, but one can easily construct them by placing the helices on the polyhedral ribs in all possible ways such that each vertex corresponds to one end of a helix (Murzin and Finkelstein, 1988)]. The packings with inter-helical angles favorable for close helix contacts (see Fig. 14.9) are observed in proteins more often than others. *((A) Adapted from Murzin, A.G., Finkelstein, A.V., 1988. General architecture of α-helical globule. J. Mol. Biol. 204, 749–770.)*

protein chain envelops its hydrophobic core, as if taking a continuous path along the ribs of a quasi-spherical polyhedron.

Now let us see how close packing forms in a protein globule. The existence of such packing follows from observations that protein is as compact and solid as an organic crystal although it may resemble a glass by the criterion of the free volume distribution (Liang and Dill, 2001). However, it is still to be explained how it comes about that such packing emerges regardless of the vast variety of most intricate shapes of side-groups of a protein chain.

Actually, the outline of close-packing formation is more or less clear only for α-helices (Crick, 1953; Efimov, 1977, 1979; Chothia et al., 1977, 1981) [for β-sheets, it is more intricate due to lesser stiffness of β-strands, see Finkelstein and Nakamura (1993)], and that is why it would not be out of place to consider it here.

The first model of the close packing of α-helices, that of the "knob (side chain) into hole (between side chains)" type, was proposed by Crick (1953) earlier than the solution of the first 3D protein structure. Later, this model was further developed by Efimov (1977, 1979), and independently, by the Chothia-Levitt-Richardson team (1977, 1981), and by now it has acquired the "ridge (of side chains) into groove (between them)" description.

According to this model, side chains in the surface of an α-helix tend to turn about the knobs formed by C^{β} atoms and form ridges separated by grooves. The "ridges and grooves" prove to be a bit better in describing the reality than "knobs and holes" because a turn of one "extended knob" (one side chain) towards another (another side chain) can make this or that "ridge composed of knobs" more definite. There are two types of ridges (and their parallel grooves): those of the "+4" type formed by side chains of residues at sequence positions "i," "$i+4$," "$i+8$," etc. (in other words, separated in sequence by four chain residues), and ridges of the "+3" type formed by side chains of residues at sequence positions "i," "$i+3$," "$i+6$," etc. (ie, separated by three residues). Fig. 14.8 shows that ridges of these two types form angles of opposite signs with the helix axis.

The close packing brings the ridges from one helix into the grooves from another.

This gives two types of possible packing (Fig. 14.9).

In the first type, "+4" ridges of one helix fit into grooves between similar "+4" ridges of the other (Fig. 14.9A; as seen, the close packing results from superimposing the overturned helix α2 onto helix α1 and turning it further until the "+4" ridges of both helices become parallel). In such packing the angle between the helix axes is close to −50 degrees. This is the most typical angle formed by helices in α-helical globules (Chothia et al., 1977; Murzin and Finkelstein, 1988). Also, it is typical for α/β and α+β proteins to be discussed later. Such an angle provides for a twist of the α-helix layer (in which the twist angle is close to −50 degrees/10 Å, where −50 degrees is the angle between axes of adjacent helices, and 10 Å is the width of an α-helix), which is in good agreement with the typical twist of a β-sheet (characterized by the same −25

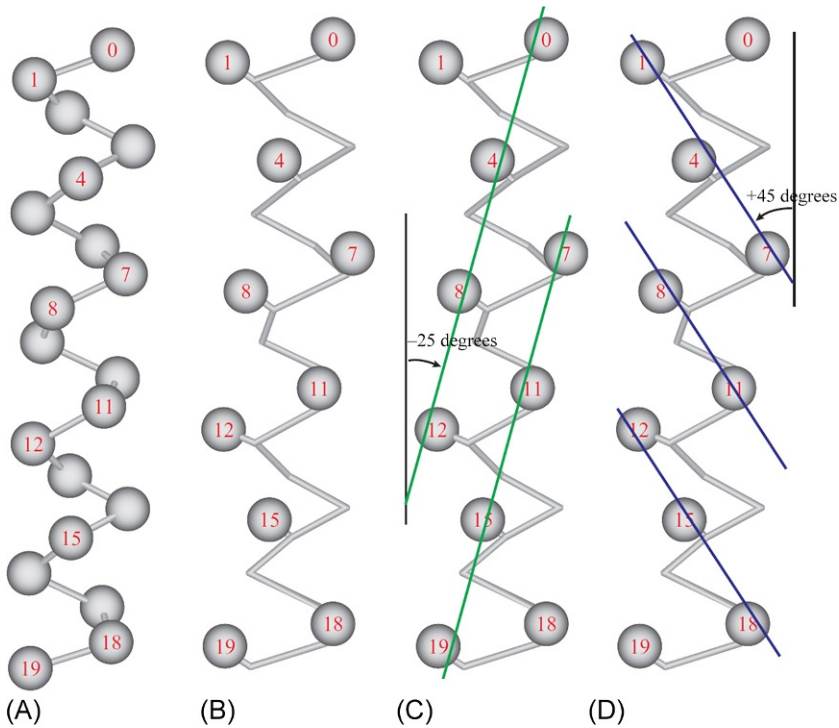


FIG. 14.8 Ridges at the surface of the α-helix. The C^{α}-atoms (A) and C^{β}-atoms (B–D) are shown. Numbered residues face the viewer. Two kinds of ridges (thin lines in the helix face) from close side-groups are shown (C, D). The ridges "+4" from side-groups "i" – "$i+4$" – $i+8\cdots$ are inclined at −25 degrees to the helix axis (C), the ridges "+3" from groups "i" – "$i+3$" – $i+6\cdots$ are inclined at +45 degrees (D); in the drawing these angles look smaller because typical ridges pass through massive side-groups, while in (C) and (D) they run through the centers of the Cβ-atoms. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure, second ed. Garland Publishers, New York (Chapter 3).)*

degrees/5 Å value, where −25 degrees is the angle between axes of adjacent β-strands and 5 Å is the width of a β-strand).

In the second type, "+3" ridges of one helix fit into grooves between "+4" ridges of the other (Fig. 14.9B). In such packing, the angle between the helix axes is close to +20 degrees. This is the most typical angle for helix contacts in long bundles that occur in α-helical globules, as well as in fibrous and membrane proteins (Chothia et al., 1977, 1981; Murzin and Finkelstein, 1988).

In addition, "+3" ridges of one helix can fit into grooves between similar "+3" ridges of the other, thereby forming an extremely short contact of almost perpendicular helices (Efimov, 1979). The contact is so short that I did not show it in Fig. 14.9 although it is quite typical of α-helical globules.

Concluding the consideration of close packing, I would like you to note that the actual interhelical angles in proteins may differ from the above given "ideal" values because side chains vary considerably in size. For the same reason, the picture of

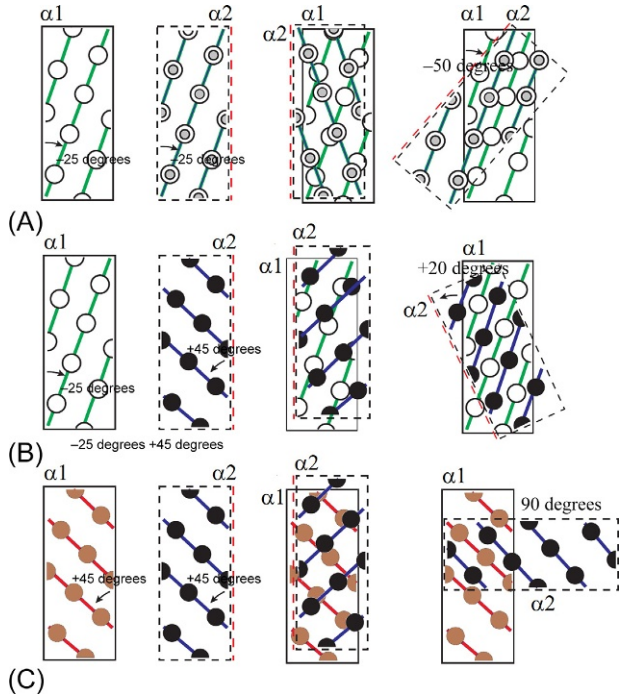


FIG. 14.9 Three variants of close packing of side chains of two helices: large contacts with helix axes inclined at −50 degrees (A) and +20 degrees (B) and small contact at 90 degrees (C). We look at the contact area through one helix (through α2 turned over through 180 degrees around its axis). The residues of the "lower" (α1) helix are shown as lighter circles and those of the upper helix (α2) by darker circles. *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure, second ed. Garland Publishers, New York (Chapter 3).)*

the ridge-into-groove fitting is slurred over in β-structures (where the side chains project much less: unlike the cylindrical α-helix, the β-sheet has a rather flat surface) and is clearly observed only in rare cases (Finkelstein and Nakamura, 1993).

Finally in this section, let us see how the close packing of helices conforms to the positioning of helices on the ribs of quasi-spherical polyhedra. As a matter of fact, it does so quite curiously. The "polyhedral" packings with angles between helices close to −50 degrees and/or +20 degrees (which are favorable for close packing of helices) are observed frequently; others are rare (Murzin and Finkelstein, 1988) although occasionally these can be observed too. For example, the first of the two three-helix packings shown in Fig. 14.7, a bundle with a right-handed twist, causes inter-helical angles of −60 degrees (close to the −50 degrees angle optimal for close packing, see Fig. 14.9A). This three-helix bundle is observed frequently. The other packing, with the left-handed twist, causes interhelical angles of +60 degrees (which differs greatly from

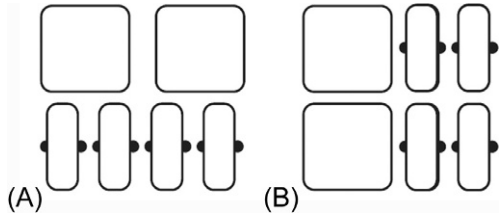


FIG. 14.10 (A) The layer structure of mixed (α/β and α+β) proteins viewed along the α-helices and β-strands to stress their close packing (helix ends are shown as squares and strand ends as rectangles). (B) α-Helices and β-strands cannot belong to the same layer because this would cause dehydration of H-bonds at the β-sheet edge (H-bond donors and acceptors in the β-sheet are shown as dots).

all angles optimal for close contacts, ie, −50 degrees, +20 degrees, and 90 degrees), and this bundle can be observed an order of magnitude less frequently.

Now let us consider "mixed" proteins built up from β-sheets and α-helices. Typically, they consist of separate α- and β-layers, and never have "mixed" ones, which would cause energetically unfavorable dehydration of H-bonds at the β-sheet edges (Fig. 14.10) (Finkelstein and Ptitsyn, 1987).

There are α/β (α slash β) and α+β (α plus β) proteins (or rather, domains), and sometimes they are combined into the common class of α&β (ie, "α and β") or α-β proteins.

In α/β domains the β-structure is parallel, the α-helices are also parallel to one another (and antiparallel to the β-strands), and along the chain they alternate (Levitt and Chothia, 1976):

Two folding patterns are most typical for α/β proteins (Fig. 14.11): the α/β-cylinder where the β-cylinder lies inside a cylinder formed by α-helices (Richardson, 1977, 1981) and the "Rossmann fold" (Rao and Rossmann, 1973) where a more or less flat (except for the ordinary propeller twist) β-layer is sandwiched between two α-helix layers whose twist is complementary to that of the β-layer.

Unlike previously considered domains, α/β domains usually have two hydrophobic cores: in the Rossmann fold, they are between the β-sheet and each α-layer; in the α/β-cylinder the smaller core is inside the β-cylinder, while the larger one is between the β- and α-cylinders.

β-Cylinders are formed by more or less straight β-strands. Each pair of neighboring (H-bonded) strands has the usual propeller twist. Therefore, the

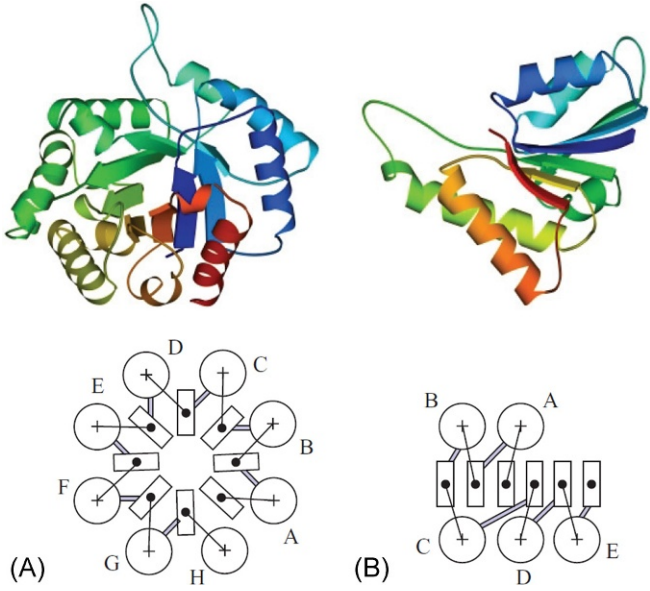


FIG. 14.11 Typical folding patterns of α/β proteins and their simplified models as viewed from the β-layer butt-end: (A) the "α/β-cylinder" in triose phosphate isomerase; (B) the "Rossmann fold" in the NAD-binding domain of malate dehydrogenase. The detailed drawing of the former shows a viewer-facing funnel formed by rosette-like loops and directed towards the center of the β-cylinder. The latter has a crevice at its upper side; the crevice is formed by loops going upwards and downwards from the β-sheet.

strands form an angle with the cylinder's axis, and the β-cylinder has a hyperbolic shape (Fig. 14.12). The β-cylinder is rigid, being stitched up with a closed hydrogen bond network. The H-bonds are perpendicular to the strands. Going from one residue to another along the line of H-bonds (along the shaded band in Fig. 14.12), one returns to the initial strand but not to the initial residue (because the strands are tilted with respect to the cylinder's axes). The resulting shear between the two ends of the hydrogen bond line is expressed as a number of residues. This number is even, since the H-bond directions alternate along the strand. Two digits, the number of β-strands and the "shear number," allow a precise discrete classification (Murzin et al., 1994) of the closed β-cylinders given by Lesk and Chothia (Lesk, 2001).

By the way, there also exist α/β "almost" cylinders that do not complete the circle and, hence, have no closed hydrophobic core inside the β-cylinder. They are known as "α/β-horseshoes" and contain up to a dozen and a half α/β repeats.

Usually, an α/β-cylinder contains eight α- and eight β-segments, and the topologies of all α/β-cylinders (and of "α/β-horseshoes", too) are alike: β- and α-segments form a right-handed superhelix, where β- and α-segments adjacent in the chain are antiparallel, while two β-strands forming a β-α-β unit are parallel and form an H-bonded contact with each other (see Figs. 14.14–14.16).

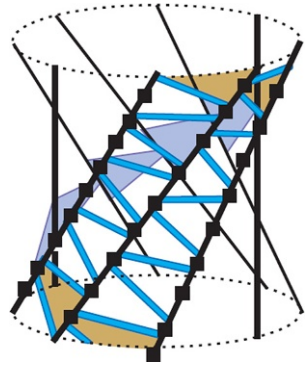


FIG. 14.12 The closed β-cylinder. H-bonds *(the blue lines)* are shown for one strand only. One line of H-bonding is shown as a shaded band. The number of β-stands is 8, and the "shear number" *(follow the shaded band!)* also is equal to 8 in the given case. *(Adapted from Lesk, A.M., 2001. Introduction to Protein Architecture. Oxford University Press, Oxford (Chapter 4).)*

Presumably this overall structure provides particular stability to the protein globule, since numerous protein globules with such architecture (10% of all proteins) display close similarity in their shape, although most of them have nothing in common as concerns their origin (ie, nothing in common as concerns their amino acid sequences) or functions.

No common functions, no similarity in the sequences or the structure of their active sites; but when it comes to *location* of the active site, α/β-cylinders have much in common: each architecture has a special place (a dent on the surface at the axis of the β-cylinder), as if specially designed for the active site, no matter what function it performs.

I would like to draw your attention to the "funnel" on the axis of the β-cylinder (Fig. 14.13A). As you can see, it is a dent in the overall protein architecture; this dent is determined by the folding pattern and is not covered with loops. This is where the active site is located. Or rather, one of two such "funnels" (to be found at both ends of the β-cylinder) is used to house the active site—it is the one where C-termini of β-strands and N-termini of α-helices are directed (Branden and Tooze, 1991; Lesk, 2001). These termini (connected with relatively short loops and having numerous open NH groups at N-ends of the helices) are believed to be most useful in binding various substrates. This is, however, still to be studied.

In the Rossmann fold the active site is located similarly: it is in a dent, in the crevice, and again in the crevice to which the C-termini of β-strands and N-termini of α-helices are directed (Rao and Rossmann, 1973). The only difference is that this crevice is formed not by loops drawn outwards from the cylinder center but by loops some of which are drawn to the upper and some to the lower α-layer (Fig. 14.13B).

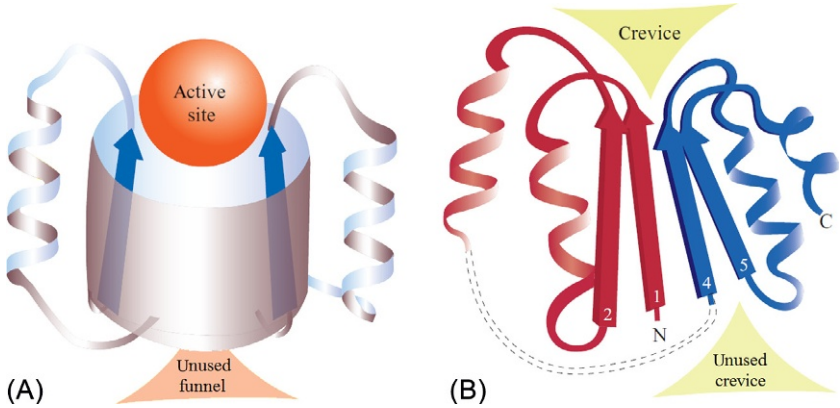


FIG. 14.13 Typical locations of the active site in α/β proteins: (A) in the "funnel" on the α/β-cylinder axis; (B) in the crevice formed by separated loops in the "Rossmann fold". *(Adapted from Branden, C., Tooze, J., 1991. Introduction to Protein Structure, second ed. Garland Publishers, New York (Chapter 4).)*

> *Inner voice*: Should I believe that the active site always occupies an obvious dent? Maybe it happens often, but far from always?
> *Lecturer*: Right. As a rule (in about 80% of instances), the active site occupies the largest dent on the globule; in turn, this dent is usually determined by the globule architecture built up by the secondary structures. Nevertheless, in many cases a search for the active site took much time—even with the spatial structure of the protein known—and was not always successful…

Now let us consider α + β proteins. They are based on the antiparallel β-structure (in contrast to α/β proteins based on the parallel β-structure) (Levitt and Chothia, 1976; Branden and Tooze, 1991; Lesk, 2001, 2010).

The α + β proteins can be divided into two subclasses. Those of the first subclass (also known as "αβ-plaits") resemble α/β proteins in that the α-layer is packed against the β-sheet. Like α/β proteins, they are characterized by a regular alternation (though distinct from that of α/β proteins) of α- and β-regions both in the chain and in space. Proteins of the other subclass ("usual" α + β proteins) have no such alternation; their α-structures are more or less separated from β-structures in the chain.

The typical alternation of α- and β-regions in the αβ-plait is either … α-β-β-α-β … or … α-β-β-β-β-α-β-β … (Fig. 14.14). Here separate α-helices are placed between β-hairpins or β-sheets composed of an even number of β-strands. The β-strands adjacent in the sequence form *anti*parallel β-sheets; and because of the even number of β-strands between α-helices (in contrast to the *odd* number of these observed in α/β proteins), as well as the general co-linearity of the strands and helices, α-helices form *anti*parallel hairpins too. The "pleated" protein structure is observed as one of the most abundant protein architectures; it is observed, in particular, among ferredoxins and … RNA-binding proteins.

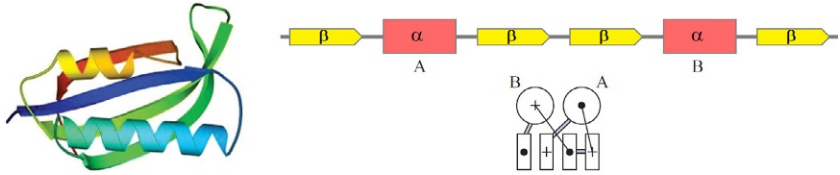


FIG. 14.14 A typical structural motif for α+β proteins: the αβ-plait in the ribosomal protein S6. The αβ-plait is distinct from other α+β proteins because it has a more regular alternation of secondary structures in the chain (in this case, the alternation is βαββαβ). S6 represents an example of the so-called "ferredoxin fold". The rainbow coloring *(blue-green-yellow-orange-red)* traces the pathway of the chain from the N- to the C-terminus. On the right, a schematic diagram of the secondary structure of this protein and its folding pattern as viewed along its almost co-linear structural elements. The helices are lettered. An α- or β-region going away from the viewer (ie, viewed from its N-terminus) is marked with "+", and that approaching the viewer with a dot.

In "normal" α+β domains (Fig. 14.15) α- and β-regions alternate irregularly and tend to form something like blocks (Levitt and Chothia, 1976). They commonly look like a β-sheet (which is often bent on itself thus forming a subdomain) covered by separate α-helices or by an α-helical subdomain. The β-structure is mostly antiparallel in α+β proteins (as in "pure" β proteins).

Now, let me draw your attention to protein topology.

A very typical feature of topology of α/β and α+β proteins (as well as of many β-proteins) is the *right*-handed (ie, counterclockwise, when approaching the viewer) topology of connections between parallel β-strands of the β-sheet (see Figs. 14.11 and 14.13–14.15) (Nagano, 1973; Richardson, 1976; Efimov, 1995). In α/β and α+β proteins such a connection usually contains

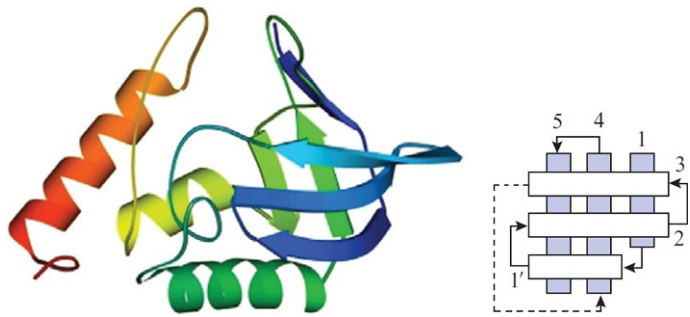


FIG. 14.15 Staphylococcus nuclease, a "normal" α+β protein characterized by a less regular (as compared with α/β proteins or αβ-plaits) alternation of secondary structures in the chain (in this case, βββαββαα); here, the α and β structures are more separated in space. The folding pattern observed in the β-sub-domain of the nuclease is called the "OB-fold" (ie, "oligonucleotide-binding fold") (Murzin, 1993). On the right: a schematic diagram of the OB-fold (the orthogonal packing of β-strands is viewed from above). The OB-fold is abundant in various multi- and mono-domain proteins. The β-strands are marked with numerals. The first strand is bent (actually, it is broken); its two halves are marked as 1 and 1′. Notice the "Russian doll effect" (Shindyalov and Bourne, 2000): one characteristic fold (the OB-fold) is a part of another characteristic fold (the nuclease fold).

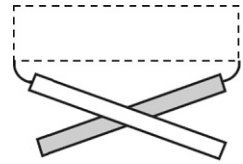


FIG. 14.16 Typical right-handed topology of connections between parallel β-strands of the same sheet. The connection usually contains an extra secondary structure.

an α-helix (Fig. 14.16). In β-proteins (and sometimes in α+β too) such a connection contains, as you may remember from the previous lecture (on "abcd" structures and so on), a β-strand from another sheet, and sometimes even a separate β-sheet with odd number of strands. It also happens, though rarely, that the connection between parallel β-strands contains neither α- nor β-structures; but in this, as well as in all other cases, it usually appears to be a *right*, not left-handed connection.

It will become clear from the next two lectures that such handedness of the connection usually contributes to protein stabilization, thus allowing a greater variety of the structure-stabilizing sequences; that's why the *right*-handed connection is quite frequently observed in different proteins, while the left-handed connection is rare.

In conclusion of this brief outline of globular protein structures, I would like to stress again that the same or very close architectures are often observed in proteins quite different both functionally and phylogenetically. This finding underlies the physical (also known in literature as "rational") classification of proteins. This will also be discussed in Lecture 15.

REFERENCES

Branden, C., Tooze, J., 1991. Introduction to Protein Structure. Garland Science, New York (Chapters 2–5, 18).

Chothia, C., 1989. Polyhedra for helical proteins. Nature 337, 204–205.

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

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

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

Efimov, A.V., 1977. Stereochemistry of the packing of alpha-spirals and beta-structure into a compact globule. Dokl. Akad. Nauk SSSR (in Russian) 235, 699–702.

Efimov, A.V., 1979. Packing of alpha-helices in globular proteins. Layer structure of globin hydrophobic cores. J. Mol. Biol. 134, 23–40.

Efimov, A.V., 1995. Structural similarity between two-layer α/β and β-proteins. J. Mol. Biol. 245, 402–415.

Finkelstein, A.V., Nakamura, H., 1993. Weak points of antiparallel β-sheets. How are they filled up in globular proteins? Protein Eng. 6, 367–372.

Finkelstein, A.V., Ptitsyn, O.B., 1987. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50, 171–190.

Lesk, A.M., 2001. Introduction to Protein Architecture. Oxford University Press, Oxford.

Lesk, A., 2010. Introduction to Protein Science: Architecture, Function, and Genomics, second ed. Oxford University Press, Oxford (Chapters 2–4).

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

Liang, J., Dill, K.A., 2001. Are proteins well-packed? Biophys. J. 81, 751–766.

Murzin, A.G., 1993. OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J. 12, 861–867.

Murzin, A.G., Finkelstein, A.V., 1983. Polyhedrons describing packing of helices in a protein globule. Biofizika (in Russian) 28, 905–911.

Murzin, A.G., Finkelstein, A.V., 1988. General architecture of α-helical globule. J. Mol. Biol. 204, 749–770.

Murzin, A.G., Lesk, A.M., Chothia, C., 1994. Principles determining the structure of β-sheet barrels in proteins. I. A theoretical analysis. II. The observed structures. J. Mol. Biol. 236, 1369–1400.

Nagano, K., 1973. Logical analysis of mechanism of protein folding. I. Prediction of helices, loops and β-structures from primary structure. J. Mol. Biol. 75, 401–420.

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

Richardson, J.S., 1976. Handedness of crossover connections in β-sheets. Proc. Natl. Acad. Sci. U. S. A. 73, 2619–2623.

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

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

Shindyalov, I.N., Bourne, P.E., 2000. An alternative view of protein fold space. Proteins 38, 247–260.

This page intentionally left blank

Lecture 15

This lecture is an attempt to explain why the majority of proteins fit a small set of common folding patterns, which should already be your impression from the previous lectures.

Actually here, we come across the "80%:20%" law. In its initial form, this law suggests that 80% of the total amount of beer is consumed by only 20% of the population.

As for the proteins, 80% of protein families are covered by only 20% of observed folds. In the previous lectures, I took the liberty of focusing mainly on these typical structures.

So, why do most proteins fit a limited set of common folds? And why not all of them (like DNA chains)? And what is behind this limited number of common folds: common ancestry? Common functions? Or the necessity to meet some general principles of folding of stable protein structures? Also, at what structural level is the similarity of proteins of distinct ancestry and function displayed?

For now we will consider these questions only qualitatively, passing to more strict answers in the next lecture, and when we know more about protein folding, I will add a couple of words on the matter.

When only a few protein structures were known (approximately up to the middle of the 1970s) each tertiary structure was believed to be absolutely unique, that is, proteins of evolutionarily different families were thought to share no similarity at all. However, with increasing information on the spatial structure of protein molecules it became more and more clear that there are "standard designs" for protein architectures (Levitt and Chothia, 1976). The architectures of newly solved proteins (or at least of their domains) more and more often appeared to resemble those of known proteins (Chothia, 1992), although their functions and amino acid sequences were utterly different. This generated the idea discussed more than once in our previous lectures, namely, that similarity of protein tertiary structures is caused *not only* by evolutionary divergence and *not* (or not only) by functional convergence of proteins, but simply by restrictions imposed on protein folds by some physical regularities (Richardson, 1977).

By the end of the 1970s it became absolutely clear that there is an intermediate structural level sandwiched between two "traditional" ones, that is, between the secondary structure of a protein and its detailed 3D atomic structure. This intermediate level is already known to us as the "folding pattern" determined by the positions of α- and/or β-regions in the globule, and it is at this level that we observe similarities in proteins having no common ancestry or function. Unlike the detailed 3D atomic protein structure, folding patterns are surprisingly simple and elegant (Fig. 15.1).

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00015-6

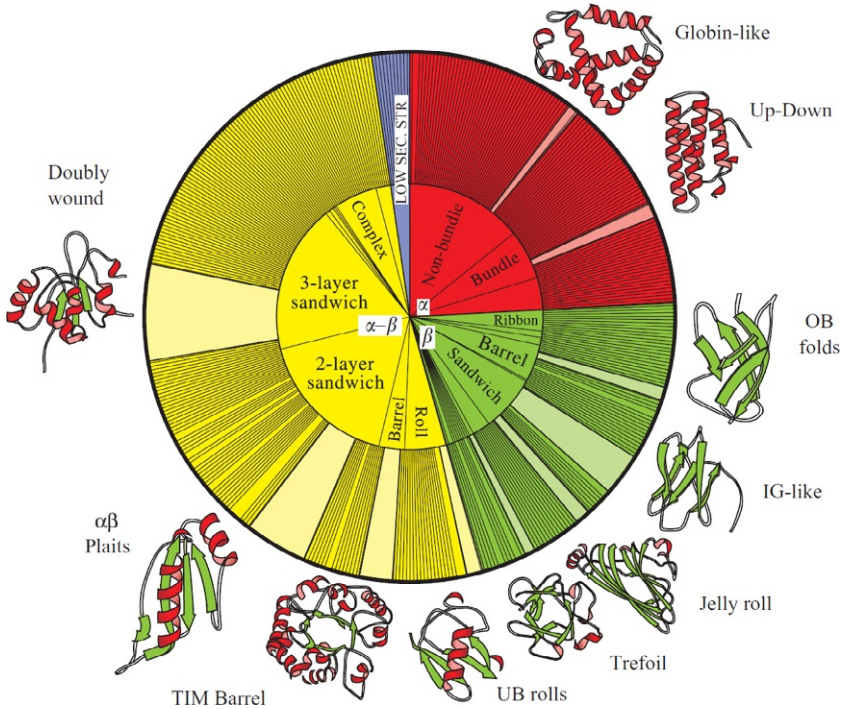


FIG. 15.1 Structural classes of proteins ("α," "β," "α-β," and "low secondary structure"), typical architectures ("nonbundle," "bundle," etc.), and typical folding patterns (topologies) according to physical classifications of proteins (CATH). Class "α-β" comprises classes that we have already considered. The sector width shows the abundance of structures of the given type in nonhomologous proteins. Note that α- and β-structures are layered, and that each layer comprises exclusively α-helices or β-strands and never houses both structures *(From Orengo et al., 1997.)*

The finding that the same or very similar architectures are often observed in proteins utterly different functionally or phylogenetically (Richardson, 1977, 1981; Ptitsyn and Finkelstein, 1980) sets the basis of physical (or "rational") classification of proteins.

The most complete computer classifications of protein folds are "Dali/FSSP," developed by Holm and Sander (1997); "CATH" (class-architecture-topology-homology) by Thornton's team (Orengo et al., 1997); and, perhaps the most popular among them, "SCOP" (structural classification of proteins) developed by Murzin et al. (1995) after he left Pushchino for Cambridge.

Actually, the classified folds refer to protein domains that are compact globules existing either separately or as a part of a multidomain protein. Classification begins (Fig. 15.1) with structural *classes* (α, β, etc.). The classes are subdivided into *architectures* of protein frameworks built up from α- and/or β-regions. In turn, the architectures are subdivided into *topologies*, that is, pathways taken by the chain through the frameworks; in other words, they are subdivided into folding patterns.

Further on, the folding patterns are subdivided into superfamilies displaying at least some sequence homology (a trace of common ancestry); those, in turn, into families with clearly displayed homology, and so on, down to the separate proteins of concrete organisms.

The physical classification of protein structures (class-architecture-topology) allows not only the systematizing of studied structures but also the prediction of protein structures yet to be found. For example, for a long time, only β-proteins composed of antiparallel β-structures were known, while a lacuna gaped where β-proteins composed of parallel β-structures should be. However, later it was filled (remember β-prisms?).

Fig. 15.2 exemplifies such a classification of the structures of protein globules, which explicitly leaves "vacancies" for possible but not (yet?) found chain folds.

Interestingly, this classification (Efimov's tree) is based on the initial small "nuclei" of various types that gradually "grow" and become more and more complicated (Ptitsyn et al., 1979; Ptitsyn and Finkel'shtein, 1979; Ptitsyn and Finkelstein 1980; Efimov, 1997). This scheme can be interpreted as an imitation of protein folding; here it should be stressed that the modern idea of this process (supported by experimental and some theoretical (Garbuzynskiy and Kondratova M.S., 2008) data to be discussed in a later lecture) is based on the concept of folding started by the formation of a small folded part of the native globule. However, these "trees" may also be interpreted as an imitation

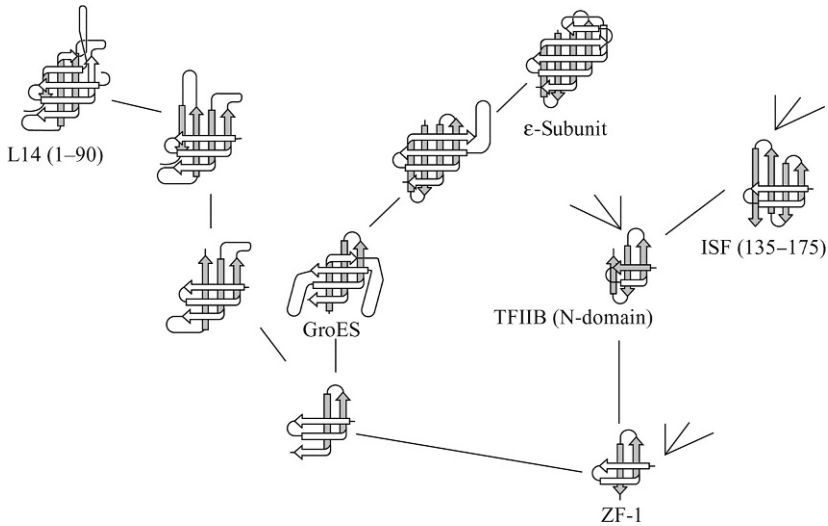


FIG. 15.2 A fragment of the "tree of structures" (Efimov, 1997) based on the initial small "nuclei" of various types gradually growing and becoming more and more complicated. Notice the "Russian doll effect," that is, that a simpler structure is contained within a more complex one. The branch fragment shown relates to β-proteins "growing" from the "β-corners" (bent β-hairpins). In this fragment, the known native structures are named, while the remaining unnamed structures were not identified in native proteins until 1997.

of the evolutionary history of proteins. The connection between these two phenomena, folding process and evolutionary history, is yet to be understood.

Now we will discuss the question that always excites a biologist: do we see the evolution of protein structures?

Actually, this question contains two questions: (1) whether we see a microscopic evolution of proteins, that is, whether we see (apart from a simple "drift," ie, change from organism to organism) some connection between a change in the protein structure and a change of the entire organism; (2) whether there is a macroscopic evolution of proteins, that is, whether their structure becomes more complex with increasing complexity of the organism.

The first question definitely has a positive answer. Although far from all the changes occurring in a protein play a clear functional role (which is stressed by Kimura's "neutral evolution theory" (Kimura, 1979)), in some cases the functional role of changes in a protein is understood and well-studied. For example, hemoglobin from a llama (a mountain animal) binds oxygen more strongly than hemoglobin from its animal relatives living on the plain. Such adaptation to living conditions is still more clearly illustrated by comparison of hemoglobins from adult animals with fetal hemoglobins: the latter has to derive oxygen from the mother organism, so oxygen binding to fetal hemoglobin must be stronger. And Max Perutz showed which microscopic changes in the hemoglobin structure are responsible for this strengthening (Perutz, 1970).

It is believed (Volkenstein, 1977; Schulz and Schirmer, 1979/2013; Cantor and Schimmel, 1980; Branden and Tooze, 1999; Lesk, 2010) that evolution often occurs through amplification of a gene with subsequent mutations of its copies, so that one copy of this gene keeps maintaining the "previous" function (and the organism's life as well), while another copy, or other copies, become free to mutate in a (random) search for a change that could adapt the protein's function to a biological need. For example, α-lactalbumin of milk undoubtedly originated from lysozyme at the advent of the vertebrates (Prager and Wilson, 1988). It is known that there is usually only one gene copy of each major protein (or rather, two identical copies, with diploidy taken into account); however, the living conditions are capable of changing the situation. An "almost fatal" dose of poison can provoke multiplication of copies of the gene responsible for its elimination (Wannarat et al., 2014). And then random mutations of these copies go into operation, then the selection…

Evolution of proteins is supported by their domain structure. It is known that the domain-encoding genes can migrate, as a whole, from one protein to another, sometimes in various combinations and sometimes individually (Lesk, 2010). Closely related domains are often observed both as parts of different proteins and as separate proteins (eg, the calcium-binding domain of calmodulin, parvalbumin, etc.; various kringle domains, and so on). Presumably, such exchange is facilitated by the intron-exon structure of genes (specifically, this is well seen in immunoglobulins) (Maki et al., 1980); however, the hypothesis that the role of

a "module" in the exchange is generally played by an exon rather than by the whole domain (Go, 1985) seems to lack confirmation.

The other question (whether there is macroscopic evolution of proteins, ie, whether their structures become more complex with increasing complexity of the organism) must most probably be answered negatively. A review of protein structures shows that the same folding patterns (specifically, those shown in Fig. 15.1) are observed both in eukaryotes, unicellular and multicellular, and in prokaryotes, although eukaryotic proteins and their domains seem to be larger than prokaryotic proteins sharing the same folding pattern, and the distribution of the most "popular" folds in eukaryotes is somewhat different from that in prokaryotes (Gerstein and Levitt, 1997; Orengo et al., 1999). Besides, eukaryotic proteins often include intrinsically disordered regions, which is not typical of prokaryotic proteins (Bogatyreva et al., 2006; Lobanov and Galzitskaya, 2012; Xue et al., 2012). However, we do not see that protein domains become more complicated with increasing organism complexity, as happens, for example, at the cellular level, as well as at the level of chromatin and organelles, down to ribosomes. (On the contrary: we see that fibrous proteins having the simplest structures, as well as "simple" disordered regions, are more typical for higher organisms rather than for prokaryotes, and especially for archaebacteria. The simplest proteins are less typical for the simplest organisms. This is strange, is it not?)

However, there exists one more important "macroscopic" structural difference, though it is not connected with the chain folds. It is as follows: the proteins of eukaryotes, of multicellular ones in particular, are much more liable to co- and posttranslational chemical modifications (such as glycosylation, iodination, etc.) (Sambucetti et al., 1986). Modification sites are marked by the primary structure, while the modification is carried out by special enzymes, and often only partially, which causes a variety of forms of the same proteins, although their biochemical activity usually remains unchanged. An alternative splicing, the privilege of eukaryotes, also contributes to the diversity of their proteins (Matlin et al., 2005).

Inner voice: Nevertheless, there is evidence that eukaryotic proteins are not only larger in size but also contain a greater number of domains than prokaryotic proteins (a typical eukaryotic protein contains four or five domains, while a prokaryotic protein only two) (see Branden and Tooze, 1999).
Lecturer: True. However, probably the general idea that eukaryotic proteins are larger in size is connected with the fact that higher organisms have many large multidomain "outer" proteins like immunoglobulins rather than with changes of the "housekeeping" cellular proteins.

It should be mentioned that the investigation of the "macroevolution" of protein chain folds is strongly hampered by the possibility of horizontal gene transfer (see Dunning Hotopp et al., 2007), which may result in penetration of "new" proteins into "old" organisms.

Inner voice: I should like to come back to the "drift" of protein structures (that you had left aside) and to a problem of the origins of protein folds. There is a hypothesis that the "jumping elements" of protein evolution are not domains, as you say, and even not a few times smaller "modules" considered by Go (1985), but short sequences of about 10 residues rather than whole domains. After all, a difference between the folds of distantly similar sequences is often caused by addition or deletion or displacement of such a short chain region (see Branden and Tooze, 1999; Fersht, 1999). And since it often contains α- or β-structure, kindred protein domains are sometimes attributed to different folding patterns. Thus, a "drift" of protein structures includes transitions from one folding pattern to another due to additions or deletions or displacements of structural modules. If so, is it possible that protein folds originated from the association of small structural modules?

Lecturer: These questions are intensively discussed from time to time, but no definite general answer has been achieved so far. The structural modules considered are so small, and their sequences are so diverse, that it is impossible to prove their relationship or reject this hypothesis. Only short fragments with similar functions (eg, some heme binding fragments) have sufficiently similar sequences. However, the same similarity is sometimes observed for "discontinuous" active sites (eg, for sites of hydrolysis) formed by remote chain residues in proteins having completely different folds. In this case, one can hardly say that these sites are "transferred" from one protein to another. Therefore, coming back to fragments that possibly serve as functional and structural modules: may be, they are transferred from one protein gene to another; maybe, they arise anew in each protein family. This is not yet known.

However, it is more correct to say that this question is still open for globular proteins. Fibrous proteins, as I have already mentioned, definitely look like multiple repeats of short fragments. Thus, their origin from "modules" looks highly probable, the more so as the repeated structural modules of fibrous proteins are often coded by separate exons. There is every reason to believe that current genomic, and especially structural genomic, and proteomics will answer all these exciting questions.

With a huge number of filed and systematized protein structures available, there arise philosophical questions such as (Finkelstein and Ptitsyn, 1987): (1) What is the physical reason for the simplicity and regularity of typical folding patterns? (2) Why are the same folding patterns shared by utterly different proteins, and what are the distinctive features of these patterns?

In scientific terms, we would like to elucidate what folding patterns are most probable in the light of the protein physics laws we have studied, how numerous these are and to what extent they coincide with folding patterns observed in native proteins. To answer these questions, we will first of all study the stability of various structures. This approach—to study stability prior to folding—is justified by the fact that the same spatial structures of proteins can be yielded by

kinetically quite different processes: in vivo (in the course of protein biosynthesis on ribosomes or during secretion of more or less unfolded proteins through membranes) and in vitro when the entire protein refolds from a completely unfolded state. This means that a detailed sequence of actions does not play a crucial role in protein folding.

Let us start with a simple question: why do globular proteins have the layered structure that we discussed in the previous lecture? In other words, let us try and see why the stability of a dense globule requires that the protein framework should look like a close packing of α- and β-layers, why it requires that α- and β-regions should extend from one edge of the globule to the other, and why it requires that the irregular regions should be outside the globule.

In principle, we have already discussed that. Hydrogen bonds are energetically expensive and therefore must be saturated in any stable protein structure. Hydrogen-bond donors and acceptors are present in the peptide group of any amino acid residue. They can be saturated when participating either in H-bonds to water molecules or in the formation of secondary structures. That is why only the secondary structures of a stable (if it wants to be so) globule but not the irregular loops have the right to be out of contact with water and belong to the molecule's interior, and why the elements containing free (from intramolecular H-bonds) NH- and CO-groups, that is, irregular loops, bends, edges of β-sheets, and ends of α-helices, should emerge at the surface.

For the sake of globule stability, extended α- and β-structures must closely surround the hydrophobic core created by their side chains, thereby screening it from water. At the same time, α-helices and β-sheets cannot share the same layer because in this case edge H-bonds of the β-sheet edges would be lost. This means that globule stability demands the formation of purely α-layers and, separately, purely β-layers (Fig. 15.1). In other words, separate α- and β-layers are stable elements of the globular protein structure, while α- and β-structures mixed in the same sheet would be a *structural defect*, or, more accurately, an *energy defect* of the protein globule. It is also evident that a *stable* globule must contain a majority of stable elements (as you know, "*what's good for General Motors is good for America*") and avoid structural defects. Since we can observe only stable globules (unstable ones fall apart and therefore cannot be observed), the observed protein structural elements must be mostly stable, and defects must be only occasionally observed.

In particular, this is true for α- and β-layers. They are stable if not mixed. And as we have seen, such layers (usually they are not flat but twisted, cylindrical, and even quasi-spherical, as in α-helical globules) are indeed typical of protein globules. The layered structure simplifies protein construction, and the large majority of domains can be represented by two-, three- or four- (rarely) layer packings.

Some proteins (especially those containing metalorganic complexes or numerous S—S-bonds) are sometimes observed to deviate from the "layered packing" scheme (Fig. 15.3), but such deviating proteins are very rare and they have "unusual" amino acid sequences as a rule.

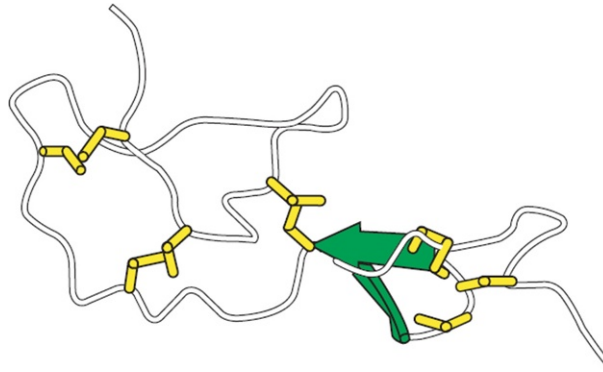


FIG. 15.3 An unusual globule with no α- and almost no β-structure (the protein huristasin, a representative of the "low secondary structure" class). This protein has a very special sequence with many Cys residues that form S—S bonds (their side chains are shown as yellow rods).

Domains with more than four layers are extremely rare, and in principle, it is clear why. They would contain too many residue positions screened from water, which means (for the 1:1 ratio of polar and nonpolar side chains typical of globular water-soluble proteins) that many polar residues would be brought into the interior of the globule. (By the way, the 1:1 ratio of polar to nonpolar residues is just what can be expected from the genetic code (Volkenstein, 1981).) This is energetically most unfavorable, and such a protein would be unstable. That is why very large (and hence, many-layered) compact globules of a "normal" amino acid composition must be unstable, and therefore, large proteins have to be divided into the subglobules that we now know as domains (Bresler and Talmud, 1944a,b; Fisher, 1964).

Actually, a chain consisting mostly of hydrophobic amino acids could pack into a very large stable globule, but such sequences are many times less numerous that those of the mixed "hydrophobic/hydrophilic" type, and, moreover, such a chain would be brought into the membrane instead of acting as a "water-soluble globular protein."

In principle, a sequence can be suggested in which some specially positioned polar side chains would provide the "cure" for all "defects," for example, for all broken hydrogen bonds between the main chain and water molecules that result from immersion of a loop or the β-sheet edge in the interior of the globule. Or a sequence can possibly be proposed that would compensate for the broken bonds with some powerful interactions, for example, with covalent (Cys-Cys) or coordinate (through the metal ion) bonds. *In principle*, this seems to be possible. But these sequences would be *very special*, and hence, *very rare*…

Perhaps this is the heart of the matter: maybe, "common" globular proteins are formed by "normal" (not too strictly selected) sequences rather than by those "strictly selected," which, therefore, are simply very rare.

Let us consider the primary structures of proteins (Fig. 15.4). Statistical analysis shows that the sequences of water-soluble globular proteins appear to be quite "random" (Finkel'shtein, 1972; Poroikov et al., 1976).

That is, in these sequences various residues are as mixed as would be expected for the result of random copolymerization. Certainly, each sequence does not result from random biosynthesis but is gene-encoded. Still, the sequences of water-soluble globular proteins look like "random" ones: they *lack the blocks* typical of membrane proteins (where clearly hydrophobic regions that must stay within the membrane alternate with more hydrophilic ones that have to form loops and even domains projecting from the membrane), and also they *lack the periodicity* characteristic of fibrous proteins (with their huge regular secondary structures).

> *Inner voice*: I cannot but note that a coded message may also look like a random sequence of letters, although this is not at all the case…
> *Lecturer*: Of course, the amino acid sequences of globular proteins are not truly random (that would imply that any sequence can fold into a globular protein; anyway, this will be discussed in later in Lecture 16). Protein sequences are certainly selected to create stable protein globules. But the shape of these globules may vary greatly. Therefore, the set of observed primary structures includes the entire spectrum of regularities inherent to all these shapes, that is, a vast set of various "codes." And when calling the primary structure "random" (or rather, "quasi-random"), we mean only that in the totality of primary structures of globular proteins, the traces of selection of protein-forming sequences are not seen as clearly (and therefore are not as restrictive) as traces of selection for periodicity in fibrous proteins or traces of selection for blocking in membrane proteins. This is what is meant when I say that amino acid sequences of water-soluble globular proteins look like random sequences.

And what is it like "to look like a *random* sequence"? This means to look like the *majority* of all possible sequences. Then in considering water-soluble globular proteins it would certainly not be pointless to try to find out which spatial

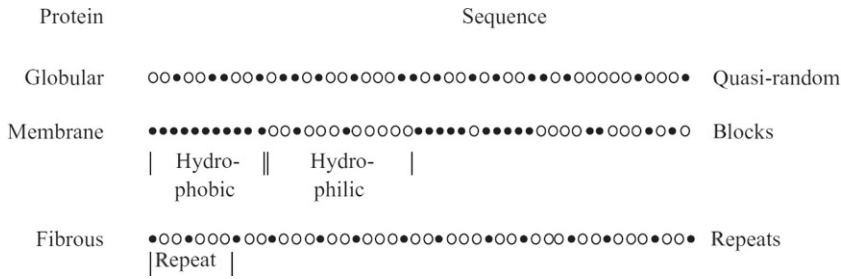


FIG. 15.4 Typical patterns of alternation of hydrophobic (•) and polar (○) amino acid residues in the primary structures of water-soluble globular proteins, membrane proteins and fibrous proteins.

structures are usually stabilized by the most common, random sequences (see Bresler and Talmud, 1944a,b; Ptitsyn, 1984; Ptitsyn and Volkenstein, 1986) or by those similar to them (by "quasi-random" sequences).

Still more. If a protein globule has a "structural defect" (eg, immersion of an irregular loop or the edge of a β-sheet in the hydrophobic core), then its stability can be ensured only by an extremely thorough selection of the amino acid sequence (to collect as many structure-supporting interactions as possible). The greater the "defect," the more rigorous the selection. And if there is no defect, then less rigorous selection is required. In other words, a "defect-free" structure can be stabilized by many sequences, a structure with a minor defect by a few, and a structure with a great defect can be stabilized only by a vanishingly small number of sequences.

And (if only physics is taken into account), the structures coded by many sequences must be observed quite often, while those coded by a small number of sequences only rarely. This is how the "physical selection" of protein structures can occur.

Since typical packings, "stacks," of secondary structures of globular proteins (Fig. 15.1) look like stable packings of random or almost random sequences should look, it is evident that, at least at the packing level, the observed result of natural (biological) selection of packings does not conflict with physical selection.

Let us proceed by considering the *folding patterns* of protein chains from the same viewpoint, from the viewpoint of structural defects and the physical selection of structures that are "defect-free" (and hence, "eligible" for many sequences).

As we have seen, protein folding patterns are often most elegant. The pathway of protein chains often resembles the patterns on pottery ornaments (Fig. 15.5). And according to the neat idea of Jane Richardson who discovered this resemblance, this is not a coincidence: both the ornament line and the protein chain aim to solve *the same* problem, that is, how to envelop a volume (in protein, its hydrophobic core) by a line avoiding self-intersection.

In proteins, this effect is achieved by surrounding the core (or two cores, as typical for α/β proteins) with secondary structures and with loops sliding over the core surface. It is also important that the loops connect *anti*parallel (and *not* parallel) secondary structures (Fig. 15.6A), and that they *do not intercross* (Fig. 15.6B). Note that the latter decreases the probability of knot formation in the protein chain.

Why are parallel connections worse than antiparallel for the secondary structure elements adjacent in the chain? Is it perhaps because then a too long irregular loop (unsupported by H-bonds) is required? Or is it because the rather rigid polypeptide chain has then to be bent, which is energetically expensive (or rather, causes a free-energy loss; see Landau and Lifshitz (1980))?

And what is wrong with loop crossing? After all, we do not mean that one loop runs into another, we only mean that one loop passes over another. Perhaps

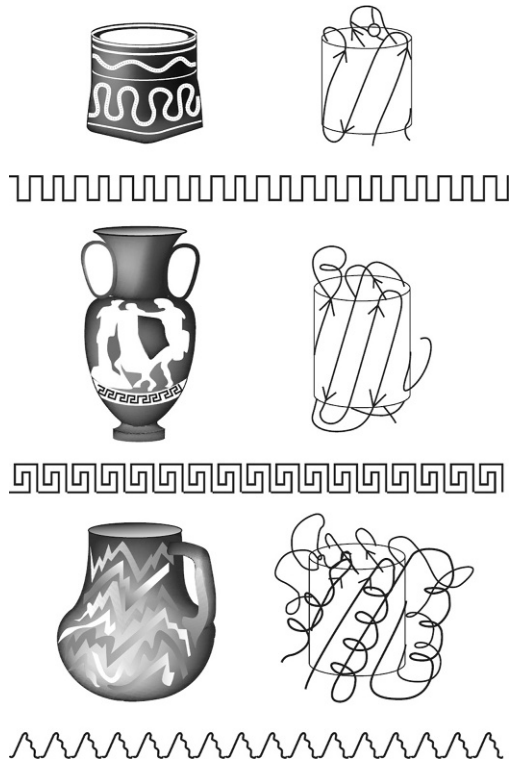


FIG. 15.5 Folding patterns of protein chains and ornaments on American-Indian and Greek pottery: two solutions to the problem of enveloping a volume with a non-self-intersecting line. On top, the meander motif; in the center, the Greek key motif; at the bottom, the zigzag "lightning" motif. *(Reprinted with permission from Richardson, J.S., 1977. β-Sheet topology and the relatedness of proteins. Nature 268, 495–500, © 1977, Macmillan Magazines Limited.)*

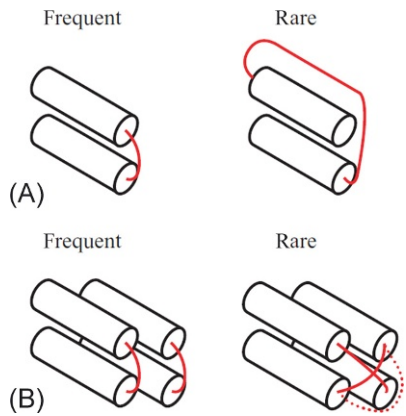


FIG. 15.6 As a rule, loops connect *anti*parallel (and *not* parallel) adjacent regions of a secondary structure (A), and any loop crossing is rarely observed in proteins, no matter if one loop covers another or by-passes it (B).

the problem is that the "lower" loop is pressed to the core and loses some of its hydrogen bonds to water molecules. And to compensate for this loss (the "energy defect"), again a "rare" sequence is needed…

It should be noted that structures with relatively small defects such as loop crossing are still observed in proteins (unlike structures with the large packing defects like mixed layers of α and β structures discussed earlier). But "faulty" protein structures are rare, which is especially significant because a "structure with a defect" can be formed in many more ways than a "defect-free" one.

Thus, we see that "structural defects" have a great effect on the occurrence of various protein structures in nature (Finkelstein and Ptitsyn, 1987).

Here, however, we might be confused by the fact that only one or two H-bonds are lost by loop crossing, that is, the loss is energetically inexpensive and amounts to only 3–5 kcal mol^{-1}. This is very much less than the total energy of interactions within the globule, which amounts to hundreds of kilocalories (as follows from protein melting data that we will consider in later lectures). Moreover, it is considerably less than the usual "margin of stability" of the native globule, that is, the free-energy difference between the folded and unfolded protein, which amounts to about 10 kcal mol^{-1} under native conditions (according to the same data). Then, why does an "energy defect" of only 5 kcal mol^{-1} virtually rule out loop crossing in native protein globules?

And one more question: why cannot the upper loop make an additional bend (dashed line in Fig. 15.6) to avoid the crossing (ie, to have it replaced by bypassing)? Perhaps the matter is, again, that the polypeptide chain is rigid and an additional bend would cost, as estimated (Finkelstein and Ptitsyn, 1987), a few (again a few!) kilocalories?

Let us postpone answering these questions until the next lecture, and consider now another characteristic feature of protein architectures, namely, that connections between parallel β-strands are almost always *right*-handed and not *left*-handed (Fig. 15.7).

In this case, the stability criterion allows us to point out the "better" of these two asymmetrical connections. Their difference is based on asymmetry of native amino acids. It causes, as you remember, a predominantly *right*-handed twist (shown in Fig. 15.7) of β-layers composed of L-amino acids.

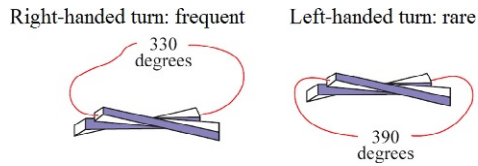


FIG. 15.7 The right-handed turn (counter clockwise rotation when the chain is approaching) of connections between parallel β-strands is frequently observed in proteins, while the left-handed turn is very rare. (The connections shown here as simple lines usually incorporate α- or β-regions.)

The angle between adjacent β-strands is close to 30 degree, such that the total rotation angle is close to 330 degree for a right-handed connection, while it is close to 390 degree for a left-handed one. As a result of polypeptide chain rigidity, the right-handed connection is more favorable: its elastic free energy is lower, although—again—by a couple of kilocalories only (Finkelstein and Ptitsyn, 1987).

Inner voice: I cannot but note that polymer elasticity is *not* an energy but rather an entropy effect (see Birshtein and Ptitsyn, 1966; Flory, 1969; Landau and Lifshitz, 1980). That is, a significantly bent chain is not as free in its fluctuation as an extended or slightly bent chain. In other words, a smaller number of conformations are possible for a bent chain than for a straight chain. Thus, you must be speaking about a fluctuating loop, whereas in the protein globule loops are *not* fluctuating, they are fixed, and no matter what pathway they take, they have a single conformation. Then, what do the entropy losses have to do with the native protein structure where the entropy of fixed chains is equal to zero anyway!?

Lecturer: I see you know polymer physics! Let me now answer your question concerning "entropic," to all appearances, defects of too sharp turns of the chain qualitatively.

The thing is that a chain with a limited choice of conformations cannot adjust well to its constituent amino acid residues. A certain conformation has low energy only for a certain sequence (or a small number of sequences) and high energy for others, that is, for most sequences. But if the chain can choose among a large set of conformations, then many more sequences can be adequately fitted. Thus, we see that the "entropy defect" can be translated into "energy defect" (or into "decreased number of sequences") language. The entropy effects will be considered more rigorously later on.

We will also consider the postponed question as to why a "defect" of only a few kilocalories per mole plays a significant role in the occurrence of protein structures. We will remember these two problems and consider them in detail in our Lecture 16 (meanwhile, you can take a look at Finkelstein and Ptitsyn (1987) and Finkelstein et al. (1995a,b)).

And right now, not to find the final answer to the latter question but just to drop a hint, let us consider relationships between the energy and other statistical rules known for protein structures. For example, let us consider the immersion of hydrophobic and hydrophilic side-groups in the protein globule, different angles of rotation in the chain (Pohl, 1971) and others. Similar to the previously discussed folding patterns, here "defects" are rare and "good elements" are usual. However, here more detailed statistics allow not only qualitative but also quantitative estimates to be made.

And the estimates obtained show the following. Both the rare occurrence of "defects" (no matter whether it is a "bad" rotation angle or the deep immersion of a polar group in the globule) and the frequent occurrence of "good elements"

(eg, salt bridges formed by oppositely charged groups) are described by the common phenomenological formula (Pohl, 1971; Finkelstein et al., 1995a,b):

$$\text{Occurrence} \sim \exp(-\text{free energy of element}/kT_C), \tag{15.1}$$

where T_C is a temperature (called "conformational" or "selective" temperature), which is close to either room (comfortable for proteins) temperature or the characteristic protein melting temperature (protein statistics do not allow us to distinguish between 300 and 370 K).

Expression (15.1) describing the occurrence of protein structural elements is surprisingly similar to Boltzmann statistics *in its exponential shape*, while its physical sense is *absolutely different*. It should be remembered that Boltzmann statistics originate from the particles' wandering from one position to another and staying for a longer time in places where their energy is lower. In contrast, structural elements of the observed (native) protein globules are fixed, they do *not* appear and disappear, and they do *not* wander from one place to another.

That is, the usual Boltzmann statistics cannot be applied to the occurrence of elements of the native protein structure. And if so, why does the statistics of occurrence of these elements have such a familiar "quasi-Boltzmann" shape?

Let us again postpone a detailed answer to this question until the next lecture. Meanwhile, let us accept, as a phenomenological fact, the estimate that the defect of about kT_C, that is, of about 1 kcal mol^{-1}, decreases the occurrence of the defect-containing structures by a few times. And, as is clear to us now, this defect must decrease by a few times the number of amino acid sequences maintaining the stability of the defect-bearing protein.

Thus, our conclusions are as follows:

1. "Popular" folding patterns look so "standard," so simple and regular, because the framework of a protein structure is a compact layered packing of extended standard solid bodies (α-helices and β-strands), and their irregular connections slide over the surface of the globule, avoiding intercrossing and crossing the ends of structural segments. Physically, this arrangement is most favorable for the globule's stability because it ensures the screening of nonpolar groups from water and H-bonding of all main-chain peptide groups immersed in the compact globule.
2. The number of such "standard" stable folding patterns is not large (numbered in the hundreds, while proteins are numbered in tens of thousands); therefore, it is not surprising that some of these "common" structures are shared by proteins different in all other respects.
3. At the same time, other (defective) folding patterns are not prohibited either. They are simply rare, since only a small number of sequences can ensure their stability. The greater the defect, the lower the occurrence of such folds.
4. A "*multitude principle*" can be proposed to describe structures of domains of water-soluble globular proteins with their typical "quasi-random" amino

acid sequences. It would read as follows: *the more sequences fit the given architecture without disturbing its stability, the higher the occurrence of this architecture in native proteins*. Proposed by us in 1993 (Finkelstein et al., 1993), this principle is now better known as the principle of "designability of protein structures" (Helling et al., 2001).

REFERENCES

Birshtein, T.M., Ptitsyn, O.B., 1966. Conformations of Macromolecules. Interscience Publishers, New York (Chapters 3, 8).

Bogatyreva, N.S., Finkelstein, A.V., Galzitskaya, O.V., 2006. Trend of amino acid composition of proteins of different taxa. J. Bioinf. Comput. Biol. 4, 597–608.

Branden, C., Tooze, J., 1999. Introduction to Protein Structure, second ed. Garland Publishing, Inc., New York, London (Chapters 2, 4, 7, 17).

Bresler, S.E., Talmud, D.L., 1944a. On the nature of globular proteins. Dokl. Akad. Nauk SSSR (in Russian) 43, 326–330.

Bresler, S.E., Talmud, D.L., 1944b. Some consequences of a new hypothesis. Dokl. Akad. Nauk SSSR (in Russian) 43, 367–369.

Cantor, C.R., Schimmel, P.R., 1980. Biophysical Chemistry. W.H. Freeman & Co., New York. (Part 1, Chapter 2).

Chothia, C., 1992. Proteins. One thousand families for the molecular biologist. Nature 357, 543–544.

Dunning Hotopp, J.C., Clark, M.E., Oliveira, D.C., Foster, J.M., Fischer, P., Muñoz Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R.V., Shepard, J., Tomkins, J., Richards, S., Spiro, D.J., Ghedin, E., Slatko, B.E., Tettelin, H., Werren, J.H., 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317, 1753–1756.

Efimov, A.V., 1997. A structural tree for proteins containing 3β-corners. FEBS Lett. 407, 37–46.

Fersht, A., 1999. Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman & Co., New York.

Finkel'shteĭn, A.V., 1972. Feedback between primary and secondary the structure of globular proteins. Dokl. Akad. Nauk SSSR (in Russian) 207, 1486–1489.

Finkelstein, A.V., Ptitsyn, O.B., 1987. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50, 171–190.

Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya., 1993. Why are the same protein folds used to perform different functions? FEBS Lett. 325, 23–28.

Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya, 1995a. Boltzmann-like statistics of protein architectures. Origins and consequences. In: Biswas, B.B., Roy, S. (Eds.), Subcellular Biochemistry. Proteins: Structure, Function and Protein Engineering, vol. 24. Plenum Press, New York, pp. 1–26.

Finkelstein, A.V., Badretdinov, A. Ya, Gutin, A.M., 1995b. Why do protein architectures have Boltzmann-like statistics? Proteins 23, 142–150.

Fisher, H.F., 1964. A limiting law relating the size and shape of protein molecules to their composition. Proc. Natl. Acad. Sci. U. S. A. 51, 1285–1291.

Flory, P.J., 1969. Statistical Mechanics of Chain Molecules. Interscience Publishers, New York (Chapters 1–3).

Garbuzynskiy, S.O., Kondratova M.S., 2008. Structural features of protein folding nuclei. FEBS Lett. 582, 768–772.
Gerstein, M., Levitt, M., 1997. A structural census of the current population of protein sequences. Proc. Natl. Acad. Sci. U. S. A. 94, 11911–11916.
Go, M., 1985. Protein structures and split genes. Adv. Biophys. 19, 91–131.
Helling, R., Li, H., Mélin, R., Miller, J., Wingreen, N., Zeng, C., Tang, C., 2001. The designability of protein structures. J. Mol. Graph. Model. 19, 157–167.
Holm, L., Sander, C., 1997. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res. 25, 231–234.
Kimura, M., 1979. The neutral theory of molecular evolution. Sci. Am. 241, 98–100. 102, 108 passim.
Landau, L.D., Lifshitz, E.M., 1980. Statistical Physics, third ed. A Course of Theoretical Physics, Vol. 5. Elsevier, Amsterdam–Boston–Heidelberg–London–New York–Oxford–Paris–San Diego–San Francisco–Singapore–Sydney–Tokyo (Section 151).
Lesk, A., 2010. Introduction to Protein Science: Architecture, Function, and Genomics, second ed. Oxford University Press, Oxford, New York (Chapters 2–4, 7).
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Lobanov, M.Y., Galzitskaya, O.V., 2012. Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes. Mol. Biosyst. 8, 327–337.
Maki, R., Traunecker, A., Sakano, H., Roeder, W., Tonegawa, S., 1980. Exon shuffling generates an immunoglobulin heavy chain gene. Proc. Natl. Acad. Sci. U. S. A. 77, 2138–2142.
Matlin, A.J., Clark, F., Smith, C.W.J., 2005. Understanding alternative splicing: towards a cellular code. Nat. Rev. 6, 386–398.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH—a hierarchic classification of protein domain structures. Structure 5, 1093–1108.
Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo Conte, L., Thornton, J.M., 1999. The CATH database provides insights into protein structure/function relationships. Nucleic Acids Res. 27, 275–279.
Perutz, M.F., 1970. Stereochemistry of cooperative effects in haemoglobin. Nature 228, 726–734.
Pohl, F.M., 1971. Empirical protein energy maps. Nat. New Biol. 234, 277–279.
Poroikov, V.V., Esipova, N.G., Tumanian, V.G., 1976. Distribution of identical amino acid residues in the primary structure of proteins. Biofizika (in Russian) 21, 397–400.
Prager, E.M., Wilson, A.C., 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J. Mol. Evol. 27, 326–335.
Ptitsyn, O.B., 1984. Protein as an edited statistical copolymer. Mol. Biol. (in Russian) 18, 574–590.
Ptitsyn, O.B., Finkel'shtein, A.V., 1979. Folding and topology parallel β-structure. Biofizika (in Russian) 24, 27–31.
Ptitsyn, O.B., Finkelstein, A.V., 1980. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? Quart. Rev. Biophys. 13, 339–386.
Ptitsyn, O.B., Volkenstein, M.V., 1986. Protein structures and neutral theory of evolution. J. Biomol. Struct. Dyn. 4, 137–156.
Ptitsyn, O.B., Finkelstein, A.V., Falk, P., 1979. Principal folding pathway and topology of all-beta proteins. FEBS Lett. 101, 1–5.
Richardson, J.S., 1977. β-Sheet topology and the relatedness of proteins. Nature 268, 495–500.
Richardson, J.S., 1981. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34, 167–339.

Sambucetti, L.C., Schaber, M., Kramer, R., Crowl, R., Curran, T., 1986. The fos gene product undergoes extensive post-translational modification in eukaryotic but not in prokaryotic cells. Gene 43, 69–77.

Schulz, G.E., Schirmer, R.H., 1979/2013. Principles of Protein Structure. Springer, New York (Chapter 9).

Volkenstein, M.V., 1977. Molecular Biophysics. Academic Press, London, NY (Chapters 1, 4).

Volkenstein, M.V., 1981. Biophysics. Nauka, Moscow (Chapters 4, 6, in Russian).

Wannarat, W., Motoyama, S., Masuda, K., Kawamura, F., Inaoka, T., 2014. Tetracycline tolerance mediated by gene amplification in *Bacillus subtilis*. Microbiology 160, 2474–2480.

Xue, B., Dunker, A.K., Uversky, V.N., 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. 30, 137–149.

This page intentionally left blank

Lecture 16

We will now discuss in detail how general structural regularities are connected with protein stability and with the number of protein structure-coding amino acid sequences.

We learned that the framework of a typical protein globule is a compact packing of layers built up from extended solid bodies (α-helices and β-structures), and that irregular connections slide over the surface of the globule virtually never intercrossing or crossing the ends of structural segments (Fig. 16.1).

We concluded that the physical reason for such an arrangement is its contribution to the stability of the globule, since it provides screening of nonpolar side chains from water simultaneously with H-bonding of the main-chain peptide groups when they are immersed in the compact globule. In turn, this increased stability allows a greater number of amino acid sequences to fit the given architecture without its destruction (which can cause a more frequent occurrence of this architecture in native proteins—we called this "the multitude principle").

As we have mentioned, the peculiarity of typical water-soluble globular proteins consists in the absence of any striking features common for their primary structures (and that is why these sequences are most diverse and numerous, so the "multitude principle" can be applied to them). Indeed, their primary structures are free of any obvious correlation, such as the periodicity characteristic of fibrous proteins or block alternation typical of membrane proteins; their polar groups are rather evenly mixed with nonpolar ones. Moreover, in water-soluble globular proteins, the amounts of polar and nonpolar residues are almost equal. As a result, their primary structures look very much like "random copolymers" synthesized from hydrophobic and hydrophilic amino acids.

Are these "random" sequences compatible with the compact chain fold in the globule? (I have to note that the first to pose this question were Bresler and Talmud (1944a,b).) In particular, are they compatible with the observed secondary structures (whose share in the protein chain is somewhat above a half)? To answer these questions, let us consider hypothetical "protein chains" that result from occasional copolymerization of equal amounts of polar and nonpolar groups (Ptitsyn and Volkenstein, 1986; Finkelstein and Ptitsyn, 1987).

To be able to fit a compact globule, an α- or β-structural segment should have a continuous hydrophobic surface that includes several hydrophobic groups. An α-helical surface is formed, as we know, by nonpolar residues positioned as $i-(i+4)-\cdots$ (sometimes, as $i-(i+3)$) in the chain, while the alternation $i-(i+2)-\cdots$ is suitable for the hydrophobic surfaces of β-strands (Fig. 16.2). It can be easily shown that even a random copolymer contains

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00016-8

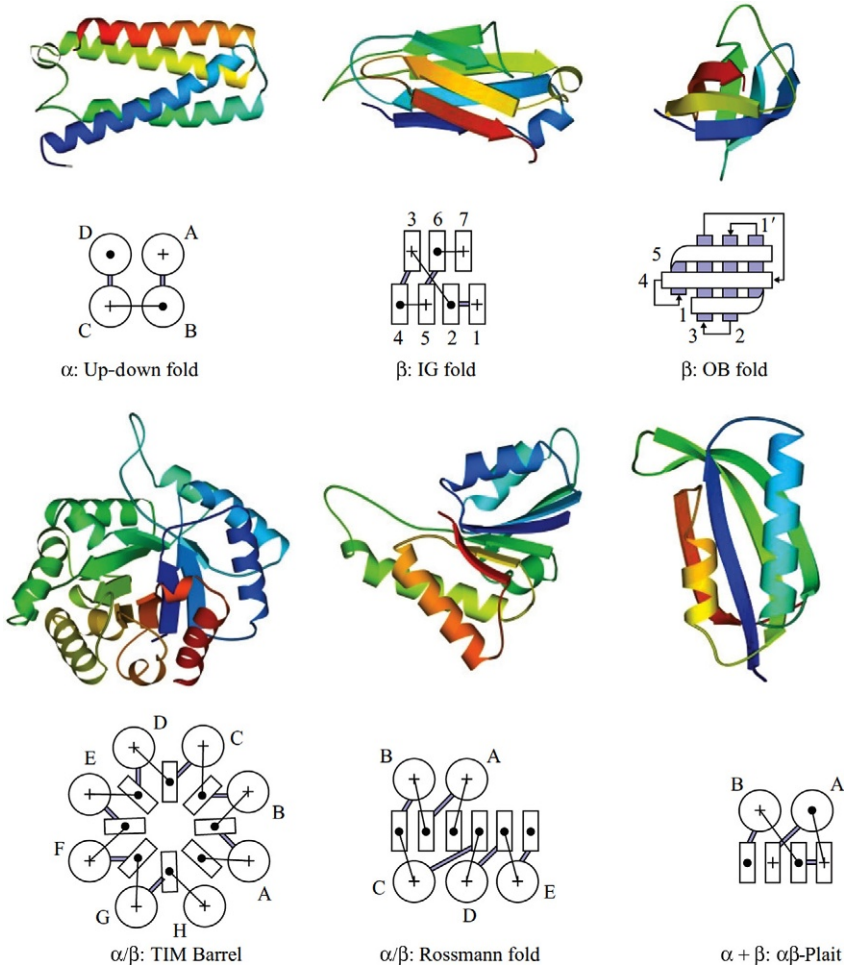


FIG. 16.1 Typical folding patterns of the protein chain in α, β, α/β, α+β, and proteins. Simplified diagrams are shown below each. The packings of α- and β-structures are layered, and each layer is composed of either α-helices or β-strands but never contains both structures.

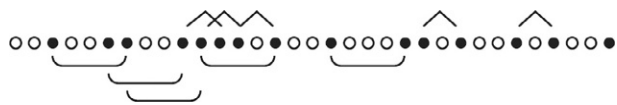


FIG. 16.2 The typical pattern of alternation of hydrophobic (●) and polar (○) amino acids in "quasi-random" primary structures of water-soluble globular proteins. Arcs *(below)* and angles *(above)* indicate the positions of potential hydrophobic surfaces suitable for α-helical and β-structural segments, that is, pairs of hydrophobic residues in positions i, $i+4$ and i, $i+2$, respectively.

enough nonpolar periodic aggregates for hydrophobic surfaces of α- and β-segments of a medium-sized protein domain.

Let "p" be the portion of nonpolar groups in a copolymer, and "$1-p$" be the portion of polar ones. Then a periodic sequence of exactly r nonpolar groups restricted at its ends by two polar residues can start at a given point with the probability:

$$W(r) = (1-p)p^r(1-p) \tag{16.1}$$

The "hydrophobic surface" of the α- and β-segment forms if $r > 1$, and the average number of groups involved is

$$\langle r \rangle = \sum\nolimits_{r>1}[W(r)\cdot r]/\sum\nolimits_{r>1}W(r) = \sum\nolimits_{r>1}[p^r \cdot r]/\sum\nolimits_{r>1}p^r = 2 + p/(1-p) \tag{16.2}$$

(I have taken the liberty of omitting the summation of series because this can be found in any maths handbook.) At $p = 1/2$, both an "average" α-helix and an "average" β-segment include $\langle r \rangle = 3$ of regularly positioned hydrophobic groups, that is, 3 ± 0.5 of the full periods of the α- or β-structure (± 0.5—because the hydrophobic group may be at the beginning, or in the middle, or at the end of the period). The expected average numbers of residues in α- and β-segments (their periods are 3.6 and 2) are $\langle n_\alpha \rangle = 11 \pm 2$ and $\langle n_\beta \rangle = 6 \pm 1$, respectively, which practically coincide with the average lengths of α- and β-segments in globular proteins (Finkelstein and Ptitsyn, 1987). Interestingly, in random sequences, as well as in primary structures of real proteins, the residue clusters good for α-surfaces often overlap those good for β-surfaces (Fig. 16.2).

Similar estimates show that the average length of loops between secondary structures in a random copolymer amounts to about $3 + 0.5p^{-2}$, that is, at $p \approx 1/2$, the loops should be somewhat shorter, on the average, than the secondary structure segments—which is indeed observed.

Thus, a random copolymer provides continuous hydrophobic surfaces that can stick α- and β-segments to the hydrophobic core at least with their one side, while the loops are relatively short. Therefore, "mediocre" random sequences are quite capable of folding into at least a two-layer arrangement of secondary structures.

Inner voice: However, one should not forget that in some, though not many, proteins (eg, in hemagglutinin or leucine zipper) there are extremely long helices that do not fit the above principles. And in some other proteins (eg, in superoxide dismutase) there are extremely long disordered loops…
Lecturer: True. These exceptions look either like blocks borrowed from fibrous helical proteins or (as concerns long loops) like anomalous hydrophilic blocks. But *on the average*, *in general*, α-, β-, and irregular segments are not too long, and their length is close to that expected for a "random" sequence containing equal proportions of hydrophobic and polar groups.

The harmony between random sequences and compact, potentially stable shapes of globules exists as long as the chain comprises fewer than ~150 residues. However, as the globule increases, as the number of its secondary structure layers grows, its "eligibility" for a random amino acid sequence decreases. This is explained by the fact that the segments belonging to the protein interior must be almost exclusively composed of hydrophobic residues, because otherwise the globule would not survive the presence of numerous water-screened hydrophilic groups and would explode, and the length of such a sequence must be proportional to the diameter of the globule. A small number of such long and almost exclusively hydrophobic segments may also be built up in a "random" copolymer from ~50% of hydrophobic and ~50% of hydrophilic residues, but only a small number indeed. Therefore, for a random sequence, we can expect not more than two or three, or sometimes four, layers of secondary structures, and this is what we really observe in single-domain water-soluble globular proteins and in the domains of such proteins (Fig. 16.1). And, for the reasons given above, large proteins must be composed of subglobules, which we know as domains, and this is indeed observed.

Now we have to return to the two questions left unanswered since the previous lecture, namely:

1. Why an "energetic defect" of a few kilocalories per mole, so very minor as compared with the total energy of protein, can virtually prohibit many protein architectures?
2. What do "entropic effects" have to do with the native protein structure where the chain is known to be fixed?

We start with the first question concerning the manifestation of the "defect" energy in protein architecture statistics. First, we will aggravate the matter: as you may remember, the stated (at the qualitative level) low occurrence of "structural defects" was supported by the observed "quasi-Boltzmann" statistics of "small elements" of protein structures, which we have to understand as well. And the statistics of "small elements" are exactly what we start with.

For example, let us consider the statistics of the distribution of amino acid residues between the interior and the surface of a protein globule and consider the interrelation between these statistics and the hydrophobicity of amino acid residues.

The hydrophobicity of amino acid residues is usually measured as the free energy of their transfer from octanol (simulating the hydrophobic core of a protein) to water. In Fig. 16.3 this free energy of transfer (divided by RT, at $T \approx 300$ K) is the vertical coordinate, while the horizontal coordinate is the logarithm of the ratio between the frequencies of the outer and inner residues in proteins. As seen, the points are arranged more or less linearly with a slope around 1–1.5.

Thus, the observed statistics of residue distribution between the interior and the surface are more or less adequately described by the expression:

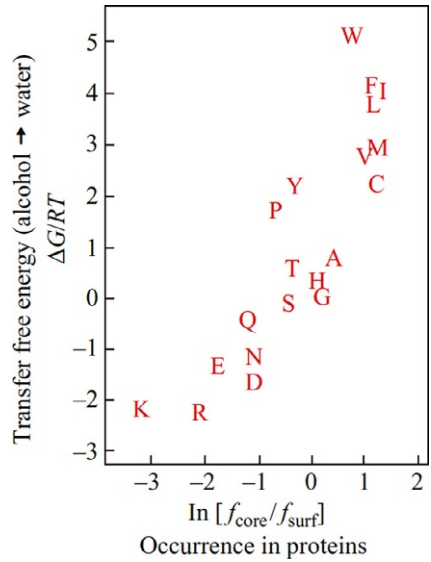


FIG. 16.3 Experimentally found free energy of transfer of residue side-groups from a nonpolar solvent to water ($\triangle G$, expressed in RT units), and "apparent free energy of transfer of a residue from the protein core to its surface," derived from observed frequencies of residue occurrence in the interior (f_{core}) and on the surface (f_{surf}) of the protein using the formula $\triangle G_{\text{app}}/RT = -\ln [f_{\text{surf}}/f_{\text{core}}]$. *(Adapted from Miller, S., Janin, J., Lesk, A.M., Chothia, C., 1987. Interior and surface of monomeric proteins. J. Mol. Biol. 196, 641–656.)*

$$\text{Occurrence} \sim \exp(-\text{free energy in given medium}/kT_{\text{C}}) \tag{16.3}$$

where the "conformational temperature" T_{C} is about 300–400 K.

Inner voice: I would be more cautious as concerns the data presented in Fig. 16.3 because in experiment, hydrophobicity was derived from residue transfer from water to high-molecular-weight alcohol. But why is the protein hydrophobic core believed to be like an alcohol? As well as the fact that alcohol is liquid while the core is solid, purely hydrophobic cyclohexane could be proposed as probably a better model of the hydrophobic core. And since the solubility of polar groups in cyclohexane is extremely low, the transfer free energy coordinate in Fig. 16.3 would inevitably be much more *extended*… True, the correlation between hydrophobicity and occurrence will also be preserved in this case, but the slope of the "cyclohexane line" drawn through the experimental points will appear to be much greater, around 3–4, or so. Incidentally, this is close to the slope, shown in the upper part of Fig. 16.3, that refers to hydrophobic amino acids which, by the way, would be least affected by replacing the core-simulating agent…

Lecturer: What is to be used to model the hydrophobic core really deserves consideration. I would say, an alcohol is still better than cyclohexane because of the presence of polar (NH and especially CO) groups in the protein core

> (although these usually participate in H-bonds within the core-surrounding secondary structure, the CO group is still capable of forming a "fork-like" H-bond, one branch of which remains unsaturated in the secondary structure). So the hydrophobicity coordinate is hardly to be adjusted to purely nonpolar cyclohexane. On the other hand, speaking of quantitative estimates, we have to bear in mind that Fig. 16.3 reflects a rough division of all side-groups into two classes (those immersed in the protein and those located at its surface), and therefore their exposures differ by only about a half of the group surface. Accordingly, the experimentally derived hydrophobicities are to be *decreased* approximately twofold, which (in contrast to alcohol replacement by cyclohexane) would decrease the tilt of the interpolation line… In principle, I do agree that the presented *numerical* data are to be treated cautiously, but the qualitative relationship between the energies of various elements and their occurrence in proteins should receive full attention.

Thus, the statistics of the occurrence of amino acid residues in the interior of a protein and at its surface exhibit a surprising *outward* similarity to Boltzmann statistics. This was first noted by Pohl (1971) for rotamers. Later, the same was shown for the statistics of many other structural elements: for occurrence of ion pairs, for occurrence of residues in secondary structures, for occurrence of cavities in proteins, and so on, and so forth. To date, this analogy has become so common that protein structure statistics are often used to estimate the free energy of a variety of interactions between amino acid residues (Miyazawa and Jernigan, 1996).

However, it should be stressed that the protein statistics resemble Boltzmann statistics only in the *exponential form* but *not* in the physical sense. As you may remember, the basis of Boltzmann statistics is that particles move from one position to another and spend more time at the position where their energy is lower, whereas in native proteins, *no* residue wanders from place to place. For example, Leu72 of sperm whale myoglobin is *always* inside the native globule and *never* on its surface. And although, according to the statistics, 80–85% of the total amount of Leu residues belong to the protein interior and 15–20% to its surface, this does not mean that each Leu spends 80–85% of the time inside the native globule and 20–15% on its surface. Rather, it means that natural selection has fixed most Leu residues at positions that belong to the interior of the globule.

That is, the usual Boltzmann statistics, the statistics of fluctuations in the usual 3D space, have nothing to do with the distribution of residues between the core and the surface of a protein. A globule has no fluctuations that could take each Leu (in accordance with its hydrophobicity) to the surface for 15–20% of the time and then take it back to the interior of the globule and keep it there for 80–85% of the time. That is, for *each separate* Leu, there is *no* Boltzmann distribution determined by its particular hydrophobicity. Then how can we explain that the occurrence of the *total amount* of leucines inside and outside proteins

agrees with the Boltzmann distribution determined from leucine's particular hydrophobicity?

Let us change the viewpoint.

Why is the predominant internal location of leucines favorable? Because it contributes to globule stability (you remember the saying "*what's good for General Motors is good for America*"). Then why were not all leucines fixed inside the protein by natural selection? Presumably, because 80–85% of internal leucines are already *enough* to ensure protein stability, and dealing with the rest would be too expensive for selection.

Let us give up the psychology of natural selection as a pointless and nonscientific topic and pursue the matter on how the internal free energy of a protein structural element affects *the number of amino acid sequences* capable of stabilizing the protein that contains the structural element in question.

For example, let us see how the Leu → Ser mutation in the protein interior can change the number of fold-stabilizing sequences.

The native (observed) structure is stable if its free energy is lower than that of the unfolded chain and of all kinds of misfolded structures. Let us now assume, for simplicity that (1) the observed fold competes only with the unfolded state rather than with other compact folds; (2) the residue's contribution to the native state stability is determined only by the residue's hydrophobicity; (3) the internal residues are completely screened from water, and the external residues are completely exposed; and (4) the residues in the unfolded protein are completely exposed to water. These statements are only approximately correct. Therefore, the following theory is rather rough—but it has the important advantage of simplicity.

The transfer free energy of a Ser side-group from the hydrophobic surroundings into water is about 0, while that of leucine is about +2 kcal mol^{-1}. Let us put aside the difference in the Leu and Ser volumes and shapes and consider only their hydrophobicity. Leu is more hydrophobic than Ser. When Leu is inside, the protein fold is more stable against unfolding than the same fold with Ser inside. This means that the sequences which can stabilize a fold with Ser inside can *also* stabilize the fold with Leu inside—but the fold with Leu inside will, *in addition*, be stabilized by some sequences which cannot stabilize the fold with Ser inside.

How does the number of fold-stabilizing sequences change when a more stable structural element (Leu inside) is replaced by a less stable one (Ser inside)? Consideration of this problem will help us to understand why occurrence of various elements depends exponentially on their free energy and what the sense of the temperature T_C in Eq. (16.3) is.

I apologize in advance for giving here calculations in a most simplified form (see Finkelstein et al. (1993, 1995a,b) and Appendix D for rigorous calculations), because a simplified form of calculations implies their incomplete accuracy. I am aiming to describe the essence of the matter without losing you in the

maths labyrinth through which one of us (A.V.F.), advised by E.I. Shakhnovich, used to ramble a lot together with A.M. Gutin and A.Ya. Badretdinov.

Let $\triangle\varepsilon + \triangle F$ be the free energy difference between the given chain fold and the unfolded state of the chain. Here $\triangle\varepsilon$ is the free energy difference for the concerned element in the fold (including all the element's interactions with the surroundings, eg, the Leu's hydrophobic free energy in the core), and $\triangle F$ is the free energy difference for the remaining chain. The fold is stable against unfolding when $\triangle F + \triangle\varepsilon < 0$, that is, when

$$\Delta F < -\varepsilon \tag{16.4}$$

The values $\triangle F$ and $\triangle\varepsilon$ depend on the amino acid sequence. Let us consider all those sequences which preserve the $\triangle\varepsilon$ value (eg, all sequences with Leu (or Ser) at the given point of the chain; for the core positions, $\triangle\varepsilon \approx 2$ kcal mol^{-1} for Leu (and ≈ 0 for Ser), while for the surface positions, $\triangle\varepsilon \approx 0$ for all residues). The value $\triangle F$ will change with the sequence. The probability P^* that $\triangle F < -\triangle\varepsilon$ is

$$P^*(\Delta F < -\varepsilon) = \int_{-\infty}^{-\varepsilon} P(\Delta F)\, d(\Delta F) \tag{16.5}$$

where $P(\triangle F)$ is the probability of the given $\triangle F$ value for a randomly taken sequence.

The $\triangle F$ value is composed of energies and entropies of many residues that independently mutate in random sequences. Therefore (according to the Central Limit Theorem of mathematical statistics), $P(\triangle F)$ has a simple, so-called Gaussian (see Fig. 16.4) form:

$$P(\Delta F) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-(\Delta F - \langle\Delta F\rangle)^2/2\sigma^2\right] \tag{16.6}$$

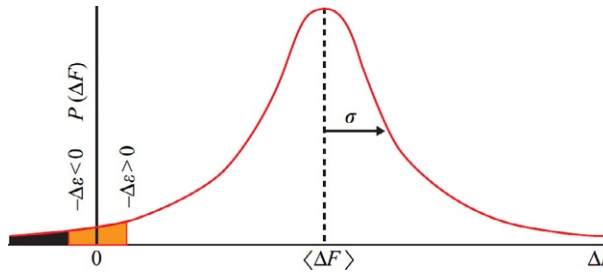


FIG. 16.4 The typical Gaussian curve for $\triangle F$ distribution among random sequences. $\triangle F$ contains the entire free energy difference between the given fold and the unfolded state of the chain, except for the fixed $\triangle\varepsilon$ value of the structural element in question. The values of $\triangle F < -\triangle\varepsilon$ (ie, those satisfying the condition that $\triangle F + \triangle\varepsilon < 0$) meet the requirements of a stable fold. The area shown in black corresponds to $\triangle F < -\triangle\varepsilon$ values at $\triangle\varepsilon > 0$, while the "red + black" area is for $\triangle\varepsilon < 0$. The latter is larger, which means that a greater number of random sequences stabilize the fold when the free energy of the element in question is below zero ($\triangle\varepsilon < 0$) as compared with its being above zero ($\triangle\varepsilon > 0$).

Here $\langle \triangle F \rangle$ is the mean (averaged over all the sequences) value of $\triangle F$, and σ is the root-mean-square deviation of $\triangle F$ from the mean $\langle \triangle F \rangle$.

I would like to remind you that the Central Limit Theorem describes the expected distribution of the sum of *many* random terms and answers the question as to the probability of this or that value of the sum. Here, "the sum of many random terms" is $\triangle F$ (composed of many interactions in a randomly taken, ie, "random," sequence). As mathematics state, for the majority ($\approx 70\%$) of sequences, the $\triangle F$ value must lie between $\langle \triangle F \rangle - \sigma$ and $\langle \triangle F \rangle + \sigma$, and the probability $P(\triangle F)$ decreases dramatically (exponentially) with increasing difference between $\triangle F$ and $\langle \triangle F \rangle$ (see Fig. 16.4).

Since the great majority of random sequences are obviously incapable of stabilizing the fold, the values of P and P^* must be far less than 1 near the "stability margin" ($\triangle F \approx 0$). Thus, the $\langle \triangle F \rangle$ value is not only positive but also *large*, much larger than σ.

The value of $(\triangle F - \langle \triangle F \rangle)^2$ is equal to $\langle \triangle F \rangle^2 - 2\langle \triangle F \rangle \triangle F + \triangle F^2$, that is, at small (as compared with large $\langle \triangle F \rangle$) values of $\triangle F$, the value $(\triangle F - \langle \triangle F \rangle)^2 \approx \langle \triangle F \rangle^2 - 2\langle \triangle F \rangle \triangle F$, and consequently,

$$P(\Delta F) \approx \left\{ \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-\langle \Delta F \rangle^2 / 2\sigma^2\right] \right\} \times \exp\left[-\Delta F \times \left(\langle \Delta F \rangle / \sigma^2\right)\right] \quad (16.7)$$

I would like you to believe (or better—to take the integral and check) that in this case the probability of $\triangle F < -\varepsilon$ is

$$P^*(\Delta F < -\varepsilon) = \int_{-\infty}^{-\varepsilon} P(\Delta F) d(\Delta F) \approx \text{const} \times \exp\left[-\frac{\Delta\varepsilon}{\sigma^2 / \langle \Delta F \rangle}\right] \quad (16.8)$$

where the constant, equal to $[(2\pi\sigma^2)^{-1/2} \times \exp(-\langle \triangle F \rangle^2 / 2\sigma^2)] \times (\sigma^2 / \langle \triangle F \rangle)$, is of no interest to us, while the really important term $\exp\left[-\left(\Delta\varepsilon / \left(\sigma^2 / \langle \Delta F \rangle\right)\right)\right]$ demonstrates that an element's free energy ($\triangle\varepsilon$) has an *exponential* impact upon the probability that a random sequence is capable of stabilizing the fold with the given element. Thus, increasing $\triangle\varepsilon$ decreases the number of fold-stabilizing sequences *exponentially*.

Note: Our previous belief was that "eligible" sequences are all those providing $\triangle F + \triangle\varepsilon < 0$, that is, at least a minimal stability of the protein structure. After a few more lectures, you will know that the "solid" protein structure should possess a certain "stability reserve." Otherwise, it would melt in our hands and be incapable of rapid and unique folding. Therefore, to be more accurate, we have to consider as "eligible" the sequences that provide $\triangle F + \triangle\varepsilon < -F_{\min}$ (where $-F_{\min} < 0$), that is, $-\triangle F < -\triangle\varepsilon - F_{\min}$. To ensure stability of the native globule, it is enough to have $F_{\min}$ of only a few kilocalories per mole, that is, it can be believed that $F_{\min} \ll \langle \triangle F \rangle$. Then Eq. (16.8) has the form:

$$P^*(\Delta F < -\varepsilon) \approx \text{const} \times \exp\left[-\frac{\Delta\varepsilon + F_{\min}}{\sigma^2 / \langle \Delta F \rangle}\right] = \text{const}^* \times \exp\left[-\frac{\Delta\varepsilon}{\sigma^2 / \langle \Delta F \rangle}\right] \quad (16.9)$$

In other words (since a change of the preexponential constant is of no interest to us), we come again to the same idea that an element's energy has an exponential impact on the probability that the element will be fixed in the protein structure as a result of the "selection for stability."

Note that a requirement of enhanced stability of the native fold (ie, an increase in the $F_{\min}$ value) decreases the number of appropriate sequences also exponentially, in proportion to $\exp\left[-\left(\Delta F_{\min}/\left(\sigma^2/\langle\Delta F\rangle\right)\right)\right]$.

The dependence (16.8) is as exponential as the Boltzmann formula but it has the term $\triangle\varepsilon$ divided not by the temperature of the environment (kT) but by the as yet unknown value of $\sigma^2/\langle\triangle F\rangle$.

What is this value? First of all, note that $\sigma^2/\langle\triangle F\rangle$ is independent of the protein size. Indeed, according to the mathematical statistics laws, the mean value (here, $\langle\triangle F\rangle$) is proportional to the number of terms summarized in $\triangle F$ (which is, in our case, approximately proportional to the protein size), while the mean square deviation from σ is proportional to the square root of the number of these terms, ie, σ^2 is also approximately proportional to the protein size.

The fact that $\sigma^2/\langle\triangle F\rangle$ *does not increase with increasing protein size* is most important: this means that the "defect's" free energy $\triangle\varepsilon$ must be compared (using Eq. 16.8) *not* with the total protein energy but rather with some characteristic energy unit $\sigma^2/\langle\triangle F\rangle$, which is something like the average energy of noncovalent interactions per residue in the chain (which is also independent of the protein size, provided the small impact of the surface is neglected).

Taken together with the exponential form of Eq. (16.8), this provides an immediate answer to the question as to why $\triangle\varepsilon$ of only a few kilocalories per mole can produce a considerable effect upon the occurrence of protein structure elements (eg, why inside a protein globule Leu residues are an order of magnitude more numerous than Ser ones). This happens because the number of fold-stabilizing sequences decreases e-fold (approximately threefold) when $\triangle\varepsilon$ increases by a value of $\sigma^2/\langle\triangle F\rangle$.

Thus, $\sigma^2/\langle\triangle F\rangle$ is something like chain energy per residue. One can present $\sigma^2/\langle\triangle F\rangle$ as kT_C, where T_C is a certain temperature (as you may remember, the "heat quantum" kT also has the sense of characteristic heat energy per particle). Exactly what temperature is T_C? For a "random" globule, there is only one characteristic temperature—that of "freezing out" its most stable fold. A protein chain has one characteristic temperature as well, that is, its denaturation temperature $\approx$350 K (which is close to that of protein's life, $\approx$300 K). Eventually, it can be shown that as long as only the minimum protein structure stability is required (ie, if selection of primary structure is determined only by this minimal restriction), the protein melting temperature is close to the freezing temperature of a random chain. Therefore, it can be suggested that the value of "conformational temperature" T_C is determined by the temperature T_M of protein melting. In other words, $\sigma^2/\langle\triangle F\rangle$ can be considered as amounting to about 0.5–1 kcal mol^{-1}.

This can be not only believed but can be shown as well. The words "can be shown" imply a quite complex theoretical physical proof, the Shakhnovich-Gutin theorem (Shakhnovich and Gutin, 1989, 1990), which I shall spare you. (Nevertheless, I would like to outline the main ideas involved. (1) Random replacement of one sequence by another changes its free energy essentially in the same way as a random replacement of one fold by another. (2) $kT_C = \sigma^2/\langle \triangle F \rangle$ is determined by the growth in the number of random sequences close to the edge separating the fold-stabilizing sequences from the others. (3) T_M is essentially determined by the growth in the number of folds close to the low-energy edge of energy spectrum of a random sequence. These two edge temperatures are close since (4) the native and denatures states of a chain are close in stability, and (5) the energy gap between the most stable (native) chain fold and its close nonnative competitors is not wide (this will be discussed in Lectures 17, 18, and in Appendix D); therefore, T_C and T_M are also close.)

Thus, we have cleared up why a defect of only a few kilocalories per mole (against the background of the much higher total energy of the protein) can virtually prohibit many motifs of protein architecture. This happens because the defect's free energy is to be compared with $kT_C \approx 0.5$–1 kcal mol^{-1} rather than with the total energy of the protein; and then we see that any defect of ≈ 1 kcal mol^{-1} decreases roughly fivefold the number of sequences "eligible" for this protein, a defect of ≈ 2 kcal mol^{-1} decreases them roughly 20-fold, and so on.

Specifically, that is why the crossing of irregular connections (this defect, caused by screening from water and loss of 1–2 H-bonds, usually costs 2–3 but never more than 5 kcal mol^{-1}) is rarely observed (Fig. 16.5) (Finkelstein and Ptitsyn, 1987).

Inner voice: I accept the argument about the loss of bonds when loops overlap, but I do not understand why the "upper" loop cannot make an additional bend and avoid overlapping the "lower" loop! Or perhaps the loops are always so short that just cannot do this?

Lecturer: No, loops are not always that short. The thing is that an additional bend of the polymer chain causes an entropy loss just like an additional bend in left-handed connection between parallel β-strands (Fig. 16.6), which was briefly discussed in the previous lecture and will be discussed in more detail right now.

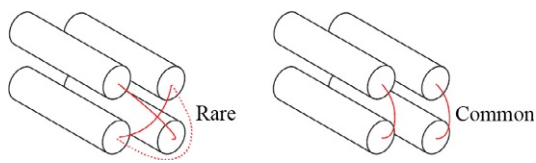


FIG. 16.5 Loop crossing (even with overlap replaced by bypassing) is rarely observed in proteins.

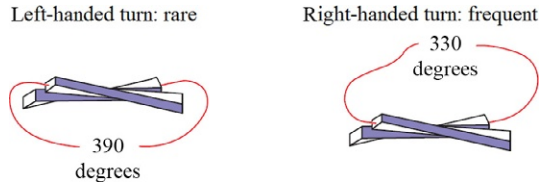


FIG. 16.6 In proteins with the right-handed (as in the figure) twist of β-sheets, the left-handed *(more bent)* connections between parallel β-strands are very rare, while right-handed ones are common.

Inner voice: OK, I will wait, but at the moment I cannot but cut in with a more general question. The entire logic of your narration is based on the assumption that sequences unable to fold into a stable structure are rejected. This is quite probable. But then, are those yielding *too stable* structures rejected too?

Actually, the observed stability of native structures is *never* very high. Moreover, a correlation between the denaturation temperature of a protein and the life temperature of the host organism has been mentioned by Alexandrov (1965). Then we have to conclude that *too* stable protein structures (and, according to your logic, *too* stable structural elements as well) are to be rejected. Seemingly, you do not take this into account at all.

Lecturer: It is difficult to make a stable protein. This has been demonstrated by the entire experience of designing *de novo* proteins. That is, sequences capable of folding into something stable are rare. It is still more difficult to create a "superstable" protein because its eligible sequences occur still more rarely. Fig. 16.4 also serves as an illustration of how small the fraction of sequences eligible for creating "superstable" protein structures (ie, those with the extremely low free energy $\triangle F$) is; see also the exponential effect of $F_{\min}$ on P^* in Eq. (16.9). The sequences good for "superstable" proteins constitute only a minor fraction of the sequences that are good for stable proteins… If the selection does not insist on "superstability" and has nothing against it, the fraction of such proteins *automatically* appears to be negligible. This explains quite satisfactorily all the experimental facts and correlations you have pointed out.

Now we, at last, pass to the question as to the role played by entropy effects in the stability of the native protein structure, although the chain there is fixed.

Let us again compare right-handed and left-handed connections between parallel β-strands (see Fig. 16.6). As we know, due to the intrinsic *right*-handed twist of a β-sheet formed by L amino acid residues, a *right*-handed connection has to turn by about 360 degrees − 30 degrees = 330 degrees, while a *left*-handed connection would have to turn by about 360 degrees + 30 degrees = 390 degrees. As stated by polymer physics, the more bent the chain, the less numerous its possible conformations. This means that the right-handed (less bent)

connection is compatible with a greater number of chain conformations than the left-handed one. This is clear. But what is the impact of the number of possible conformations on the number of sequences stabilizing this or that connection pathway?

A randomly taken sequence can provide the highest stability (among all other chain structures) of either one of the numerous conformations corresponding to the less-bent connection or one of a few conformations corresponding to the more bent connection; alternatively, it may be unable to ensure stability of any of them. The latter is certainly the most probable, since randomly composed amino acid sequences possess no stable 3D structure. However, if we have to choose only between the left- and right-handed connections, which of them has a better chance to become the more stable?

Each separate conformation containing a less bent connection is neither better not worse than each separate conformation with a more bent connection, provided that these conformations are equal in compactness, in secondary structure content, etc. Still, the less bent conformations are much more numerous…

Actually, here, we have a kind of lottery with a small chance of winning the main prize (that is creating a stable 3D structure); in this lottery, the right-handed (less bent) connection has many tickets (possible conformations), while the left-handed connection has only a few. Which of them, if either, will win the prize? Almost certainly the right-handed connection that has many more tickets and hence a better chance of success, since the probability of winning is directly proportional to the available tickets (conformations).

In other words, the broader the set of *possible* conformations, the more frequently (in direct proportion to the set range) this set contains the most stable structure of a random sequence. This is exactly what we observe in globular proteins: here the right-handed (less restricted) pathway of a connection is the rule, while the more restricted left-handed pathway (or bypassing loop with an additional bend) is the exception.

Allow me to remind you of one of these very rare exceptions. I mean the left-handed β-prism discussed a couple of lectures back. Its spatial structure is indeed unique as it consists almost entirely of left-handed connections between β-segments. And what about the twist of its β-sheet? The twist is absent… And what about its primary structure? It appears to be unique as well: first, it does not look like "random" (that would be typical of globular proteins) but contains 10 repeats of an 18-residue peptide forming each turn of the left-handed superhelix (with three β-strands per turn); second, its β-strands contain abnormally numerous glycines (which are neither L nor D amino acids and therefore do not twist β-sheet, as you know). So, a unique spatial structure is combined with a unique primary structure…

Let us consider one more problem connected with entropic defects. It concerns not folding patterns but the layered structure of the globule. Which should occur more often, proteins with α- or β-structure in the center?

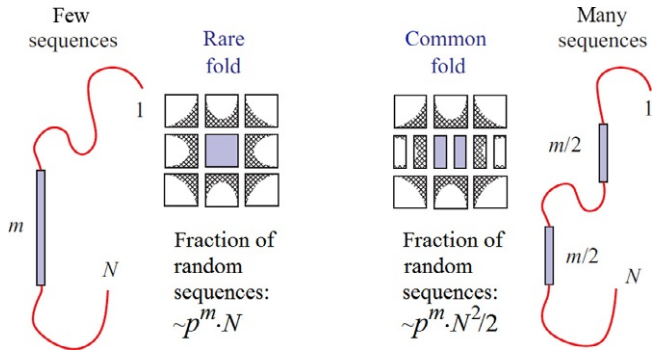


FIG. 16.7 In a large globule, a multiple-layer packing with an α-helix in its center should be less probable (and, indeed, is observed more rarely) than a multiple-layer packing with two β-segments in its center. The reason is that the segment occupying the globule's center should contain exclusively hydrophobic residues, and its large length is dictated by the large globule diameter (which is proportional to $N^{1/3}$). Since an α-helix contains twice as many residues as an extended β-strand of equal length, a central α-helix (with a large block of hydrophobic groups) will require a much less common sequence than is required for the creation of two central β-segments (with two hydrophobic blocks of half the length positioned *somewhere* in the chain). The probability of occurrence of one entirely hydrophobic block of m residues in a *given position* of the random chain is about p^m (where p is the fraction of hydrophobic residues in the chain); there is an equal probability of two hydrophobic blocks of half the length occurring in *two given positions* of a random chain: $(p^{m/2}) \times (p^{m/2}) = p^m$. However, one block can be placed in the N-residue chain in only $\approx N$ different ways, while there are many more ways ($\approx N \times N/2$) to position two blocks, that is, there are many more sequences with two short hydrophobic blocks than with one long block.

Fig. 16.7 explains why globules with β-strands in the center can be stabilized by many more primary structures than those with α-helices in the center.

And indeed, the internal α-helix occurs in proteins most rarely (the green fluorescent protein shown in Fig. 16.8 is a striking example of such the most unusual fold), while the internal β-strands are typical (eg, for "Rossmann folds," see Fig. 16.1).

The same standpoint (How often is this structural element stabilized by random sequences?) may be used to explain many other structural features observed in globular proteins. Among them, the average domain size for a chain with a given ratio between hydrophobic and hydrophilic groups (this problem was studied by Bresler and Talmud in Leningrad as far back as in 1944, and then by Fisher (1964) in the USA), and also the previously discussed average lengths of α-helices, β-segments, and irregular loops.

Thus, our analysis shows that the probability of observing the given structural element in stable folds *formed by those random sequences that are able to stabilize at least some 3D structure* should be greater, the lower the element's energy and the greater the number of conformations appropriate for this element.

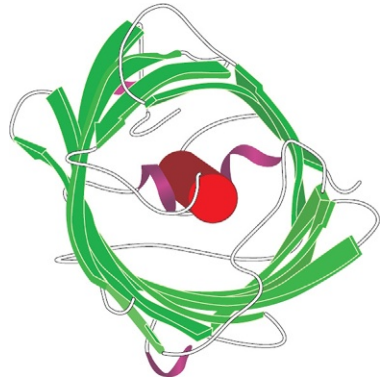


FIG. 16.8 Green fluorescent protein has an unusual fold in which the α-helix is surrounded by the β-sheet. The central α-helix is not continuous: actually, it is broken by the chromophore into two approximately equal halves (not shown).

And since the energy and the number of conformations unite in the free energy, the statistics of occurrence of various structural elements in *randomly created stable globules* should have the form:

$$\text{Occurrence} \sim \exp\left(-\text{free energy}/kT_{\text{C}}\right), \tag{16.10}$$

where T_{C} is close to (but not equal to, as a thorough analysis shows) the freezing temperature of the globule formed by a random amino acid sequence; in turn, this temperature is close to the protein melting temperature T_{M}.

And, in fact, general protein statistics have the form outlined above, which is typical for statistics of structures built up by random sequences that have been selected only for creation of stable globular structures.

Let me remind you once again of the physical basis of the above relationship. The free energy of a given structural element *exponentially* changes the number of sequences capable of stabilizing the protein whose native structure contains this element. If the element itself is stable, its host protein can stand many even unfavorable mutations, that is, the protein containing this element can be stabilized by a quite large number of sequences. If the element is unstable, its host protein demands a most thorough selection of its primary structure (stability of its 3D structure can be easily ruined by even a few mutations), and these "thoroughly selected" sequences are rare.

Here, it should be stressed that the so-called prohibited (not observed or rarely observed) protein structures are *not* impossible in principle but simply improbable because they can only be created by a small number of sequences.

Thus, structures of globular proteins resemble those which can be expected for folds of random copolymers (Ptitsyn, 1984)—or maybe for little "edited" ones (Finkelstein and Ptitsyn, 1987; Shakhnovich and Gutin, 1990; Finkelstein et al., 1993).

It appears that globular proteins could quite easily originate from random amino acid sequences (to be more accurate, from pieces of DNA coding random amino acid sequences). It would only require slight stabilizing (through a few mutations) of the most stable 3D structure of the initial random polypeptide and "grafting" an active site on its surface (to provide "biologically useful" interactions with surrounding molecules). Also, it would be necessary to clean the protein's surface of the residues that could involve it in "biologically harmful" associations (like those provoking sickle cell anemia by sticking hemoglobins together).

Inner voice: Are we to understand that you are suggesting that proteins with "new architectures" originated from random sequences rather than through strong mutations of proteins with some "old" architectures? And that the folding patterns whose stability is compatible with many random sequences originated from them many times (as many as there are homologous families in them), while other patterns covering only one homologous family each arose only once?

Lecturer: All we can say with confidence is that more than a negligibly small fraction of random sequences can give rise to proteins, and that the stability of some folds (specifically, of those often observed in proteins) is compatible with a larger number of random sequences than the stability of other, "rare" folds. As to your questions about the historical happenings, I think they cannot be answered at present. In particular, it is impossible to say whether representatives of "popular" folding patterns arose many times or not. Perhaps they originated many times from different random sequences. Perhaps only once (and not from a random sequence but from pieces of some other proteins). It is also possible that later, in the course of evolution, sequences of the same root fell so wide apart within the frames of the preserved folding pattern (since "popular" patterns are compatible with so very different sequences), that all signs of homology and genetic relationships have been wiped out, and we cannot trace them. I only want to stress that the "popular" folding patterns compatible with so many sequences give much more space for any kind of origin and subsequent evolution than the "rare" patterns.

In this connection, I cannot but note that contemporary attempts at *de novo* protein design widely use both multiple random mutations of sequences, necessarily accompanied by selection of "appropriate" (say, capable of binding to something) variants, and random shuffling of oligopeptides (accompanied by similar selection). The modern methods of selecting "appropriate" random sequences (I would like to mention perhaps the most powerful of them: it is called phage display; do read about it in molecular biology textbooks) allow

examining about 10^{12} of random sequences. The fact that "protein-like" products were now and then found among these samples shows that the fraction of such "protein-like" chains amounts to $\sim 10^{-11}$ of all random polypeptides composed of a few dozens of amino acid residues (Keefe and Szostak, 2001).

I stress that origination from random sequences is especially appropriate for globular proteins because it is their sequences that outwardly resemble "random" (ie, most abundant) copolymers (Fig. 16.2). At the same time, *non*randomness is obvious for primary (and also spatial) structures of fibrous and membrane proteins (but incidentally, the principles of their construction are simple too: repeats of short blocks for fibrous proteins and alternation of hydrophobic and hydrophilic blocks for membrane proteins).

The above analysis emphasizes the fact that evolution-yielded protein structures look very "reasonable" from the physical point of view, just like the DNA double helix and the membrane bilayer. Presumably, at the level of protein domain architectures as well, evolution does not "invent" physically unlikely structures but "selects" them from physically sound ones (ie, those that are stable and therefore capable of rapid self-organizing, as we will see soon). This is what the sense of "physical selection" consists in.

REFERENCES

Alexandrov, V.Ya., 1965. On the biological significance of the correlation between the level of thermoresistance of proteins and the environmental temperature of the species. Usp. Sovr. Biol. (in Russian) 60, 28–44.

Bresler, S.E., Talmud, D.L., 1944a. On the nature of globular proteins. Dokl. Akad. Nauk SSSR (in Russian) 43, 326–330.

Bresler, S.E., Talmud, D.L., 1944b. On the nature of globular proteins. II. Some consequences of a new hypothesis. Dokl. Akad. Nauk SSSR (in Russian) 43, 367–369.

Finkelstein, A.V., Ptitsyn, O.B., 1987. Why do globular proteins fit the limited set of folding patterns? Progr. Biophys. Mol. Biol. 50, 171–190.

Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya., 1993. Why are the same protein folds used to perform different functions? FEBS Lett. 325, 23–28.

Finkelstein, A.V., Gutin, A.M., Badretdinov, A.Ya, 1995a. Boltzmann-like statistics of protein architectures. Origins and consequences. In: Biswas, B.B., Roy, S. (Eds.), Subcellular Biochemistry. Proteins: Structure, Function and Protein Engineering, vol. 24. Plenum Press, New York, pp. 1–26.

Finkelstein, A.V., Badretdinov, A.Ya., Gutin, A.M., 1995b. Why do protein architectures have Boltzmann-like statistics? Proteins 23, 142–150.

Fisher, H.F., 1964. A limiting law relating the size and shape of protein molecules to their composition. Proc. Natl. Acad. Sci. U. S. A. 51, 1285–1291.

Keefe, A.D., Szostak, J.W., 2001. Functional proteins from a random-sequence library. Nature 410, 715–718.

Miyazawa, S., Jernigan, R.L., 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 256, 623–644.

Pohl, F.M., 1971. Empirical protein energy maps. Nat. New Biol. 234, 277–279.

Ptitsyn, O.B., 1984. Protein as an edited statistical copolymer. Mol. Biol. (in Russian) 18, 574–590.
Ptitsyn, O.B., Volkenstein, M.V., 1986. Protein structures and neutral theory of evolution. J. Biomol. Struct. Dyn. 4, 137–156.
Shakhnovich, E.I., Gutin, A.M., 1989. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys. Chem. 34, 187–199.
Shakhnovich, E.I., Gutin, A.M., 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature 346, 773–775.

Part V

Cooperative Transitions in Protein Molecules

This page intentionally left blank

Lecture 17

In a previous lecture, we considered the stability of fixed, "well-folded," "solid" protein structures. However, depending on ambient conditions, the most stable state of a protein molecule may be not solid but molten or even unfolded. Then the protein denatures and loses its native, "working" 3D structure.

Usually, protein denaturation is observed in vitro as a result of an abnormal temperature or denaturant (ie, urea, H^+ or OH^- ions (ie, abnormal pH), etc.). However, decay of the "solid" protein structure and its subsequent refolding can also occur in a living cell; specifically, this is important for transmembrane transport of proteins, for protein degradation, etc.

Moreover, even under physiological conditions, not all proteins "by themselves" have a fixed 3D structure.

Those having no fixed structure under normal cellular conditions were called "intrinsically disordered" or "natively unfolded" proteins; currently, they attract a great interest (Wright and Dyson, 1999; Uversky et al., 2000; Uversky, 2013; Dunker et al., 2001, 2008; Tompa, 2010). In some cases, the entire protein is "natively disordered," while in others, only some portions of the chain (having no X-ray-revealed structure) may be called so. Some "natively disordered" proteins are in the coil state, others are in the "molten globule" (MG) state (which will be described soon). It is now becoming clear that many protein functions require the disordered state. Native disorder is often an integral feature of proteins involved in recognition, signaling, and regulation (especially in eukaryotes), but never of enzymes (Uversky, 2013) whose work requires rigidity (we will see it in one of the next lectures).

Some of natively unfolded proteins (or natively unfolded chain segments) have a strange, simple amino acid composition, and may never have a definite spatial structure, even when contacting their targets, although many "natively unfolded" proteins acquire their unique spatial structure when binding to a ligand (Bychkova et al., 1988), or another protein, or DNA, or RNA.

It is assumed that binding of protein with natively disordered structure has that advantage that it is not too strong, and therefore, reversible even when the binding area is large (and hence, providing a significant selectivity of binding; see Schulz, 1977; Schultz and Schirmer, 1979). It is also assumed that disorder speeds up scanning of possible targets (Drobnak et al., 2013).

From my point of view, two features of intrinsically disordered proteins are of a great interest: (1) their functions (we will consider them in one of the next lectures) and (2) their ability to survive in aggressive biological environment for a rather long time. As to physics of disordered protein chains, it is not as exciting as physics of "well-folded" proteins where each atom knows its place.

Protein Physics. http://dx.doi.org/10.1016/B978-0-12-809676-5.00017-X