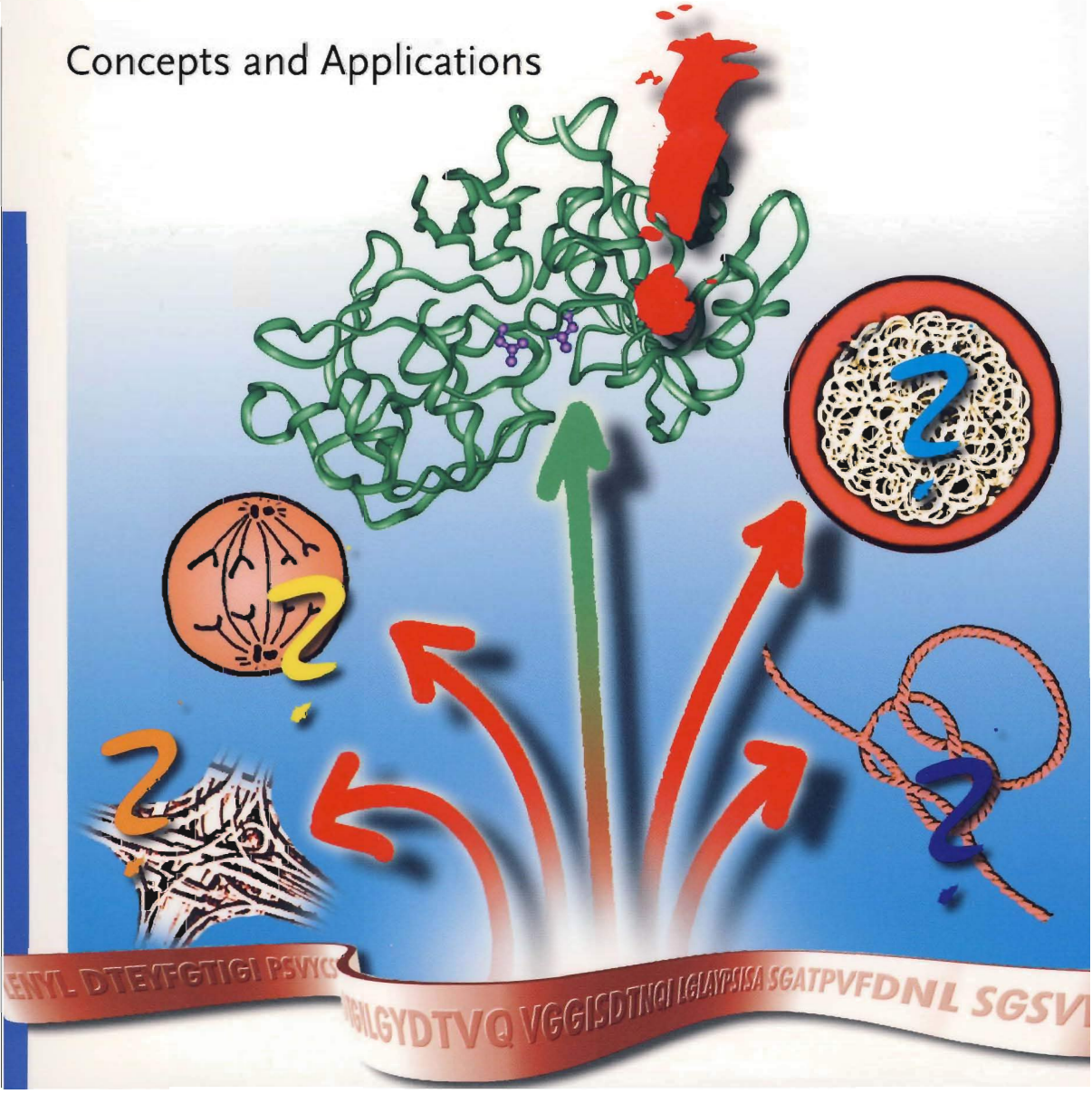
Anna Tramontano

WILEY-VCH

Protein Structure Prediction

Concepts and Applications



Related Titles

Eidhammer, I., Jonassen, I., Taylor, W. R.

Protein Bioinformatics

An Algorithmic Approach to Sequence and Structure Analysis

376 pages
2004
Hardcover
ISBN 0-470-84839-1

Höltje, H.-D., Sippl, W., Rognan, D., Folkers, G.

Molecular Modeling

Basic Principles and Applications

240 pages with 66 figures and 20 tables
2003
Softcover
ISBN 3-527-30589-0

Bourne, P. E., Weissig, H. (eds.)

Structural Bioinformatics

648 pages
2003
Softcover
ISBN 3-471-20199-5

Budis, N.

Engineering the Genetic Code

Expanding the Amino Acid Repertoire for the Design of Novel Proteins

312 pages with 76 figures and 7 tables
2006
Hardcover
ISBN 3-527-31243-9

Anna Tramontano

Protein Structure Prediction

Concepts and Applications

WILEY-VCH Verlag GmbH & Co. KGaA

Library of Congress Card No.:
applied for

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

Bibliographic information published by Die Deutsche Bibliothek
Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed biliographic data is available in the Internet at <http://dnb.ddb.de>.

Typesetting Dörr + Schiller GmbH, Stuttgart
Printing betz-druck GmbH, Darmstadt
Binding Schäffer GmbH, Grünstadt
Cover Design Grafik-Design Schulz, Fußgönheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN-13: 978-3-527-31167-5
ISBN-10: 3-527-31167-X

Dedication

This book is dedicated to two outstanding scientists who have been my guide to the fascinating field of protein structure: Professor Robert Fletterick who introduced me to the beauty of protein structure and Professor Arthur Lesk who has taught me everything I know.

Foreword

The spontaneous folding of proteins to their native states is the point at which life makes the giant leap from the one-dimensional world of DNA and protein sequences to the three-dimensional world we inhabit. Proteins must therefore have an "algorithm" by which their sequences determine their structures. This discovery, by Anfinsen almost 50 years ago, has challenged scientists to reproduce this algorithm – or at least to find another, perhaps an artificial one – to predict protein structures from their sequences. This would truly unlock the immense stores of information contained in the many genome sequences now known, to reveal the evolution and development of biological function.

Claims of progress in protein structure prediction necessitate an objective approach to evaluating them. The CASP (Critical Assessment of Structure Prediction) programmes were devised as "blind tests" of developing methods. There is consensus that CASP has stimulated, as well as recorded, the recent advances in the field. CASP came just at the time when both the growth of sequence and structure data, and the improved power of algorithms and computers, joined to make progress possible.

Anna Tramontano has here set out the current state of the art. Combining the experience of a contributor with the skills of an expositor and teacher, she has organized and presented the field. In addition to her own contributions to the development of prediction methods, she has twice served as assessor at CASP meetings. Her book sets out structure prediction in the context of the biology of the problem – protein-folding itself, of course, and the background from studies of evolution of protein sequences, structures and functions that have provided the basis for many of the most successful methods of prediction.

Several factors contribute to the clarity of the book. The illustrations are well chosen. Web links are provided. At many points of the exposition questions and answers are interpolated, as if in a lecture a member of the audience raised a hand, was recognized, and the point explained.

Protein structure prediction is at the point of maturing from an esoteric specialty to a component of the standard tools of the molecular biologist. This book will catalyse that process. The book combines a presentation of the intellectual framework of the subject, with practical aspects: If I need a model, what method should I

choose? How can I apply it? What can I expect from the result? How far can I trust it? To know how far to trust predictions, one must understand how the methods work. To contribute to the field, one must understand how to create new methods, areas where progress can be made, and how to evaluate one's own contribution (as well as those of others). This book is a good source for all these topics.

The book treats the mainline topics in protein structure prediction: Homology modelling, secondary structure prediction, fold recognition, and prediction of three-dimensional structures of proteins with novel folds. It also covers special cases, including membrane proteins and antibodies.

I and other readers must be grateful for this snapshot of state of the art at a crucial time. Of course methods will become more powerful – books like this sow the seeds of their own supersession, by fostering novel developments in the field. But this volume will remain a standard reference for the cusp of the wave.

Arthur Lesk
University Park, Pennsylvania, USA
November 2005

Table of Contents

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

Preface

The enormous increase in data availability brought about by the genomic projects is paralleled by an equally unprecedented increase in expectations for new medical, pharmacological, environmental and biotechnological discoveries. Whether or not we will be able to meet, at least partially, these expectations it depends on how well we will be able to interpret the data and translate the mono-dimensional information encrypted in the genomes into a detailed understanding of its biological meaning at the phenotypic level.

The major components of living organisms are proteins, linear polymers of amino acids, whose specific sequence is dictated by the genes of the organism. The genes, linear polymers of nucleotides, are directly translated into the linear sequence of amino acids of the encoded protein through a practically universally conserved code. This conceptually simple process is indeed biochemically rather complex, it involves several cellular machineries and is subject to an intricate network of control mechanisms. Even the problem of identifying coding regions in eukaryotic genomes is not completely solved. Far more complex is the identification of the function of the encoded proteins and this will probably be the most challenging problem for the next generations of scientists.

Proteins mediate most of the functions of an organism, and all these functions are, in general, determined by the proteins' three-dimensional structure. Natural proteins spontaneously assume a unique three-dimensional structure. Although this is not a general property of polymers, it is shared, with rare exceptions, by all natural functional proteins. It is achieved by the interplay between molecular evolution and environment-driven selective pressure. Selective pressure acts on function, function requires structure, and therefore protein sequences are selected for being able to fold into a unique structure. This peculiarity of native protein sequences is the reason for their exceptional plasticity and versatility and leads to the exquisite specificity of these macromolecules and to very precise control of their activity through complex networks of interactions. If we could understand the relationship between protein sequence, structure, and function, we could make an effective use of the large body of genetic information available for many organisms, humans included, list all their functions, and study their interplay in shaping life.

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

Protein sequence usually determines protein structure, as was established more than 50 years ago by Christian Anfinsen, the famous Nobel laureate American chemist who was the first to show that a protein can spontaneously refold to its native form and therefore that the information determining the three-dimensional structure of a protein resides in the chemistry of its amino acid sequence. Understanding the underlying rules is, however, a very challenging and as yet an unsolved problem, often referred to as the "holy grail" of biology. The problem has an enormous relevance in many fields, from medicine to biology, from biotechnology to pharmacology and therefore approximate methods for inferring the structure of a protein from its amino acid sequence are flourishing.

Such methods are an essential part of the cultural background and of the toolset of a biologist. There are several structure-prediction tools, easy to use and readily available via internet-based servers, and they are an important contribution of computational biology to the development of the life sciences. All available prediction methods have limitations, however, dictated by the hypotheses on which they rely, and this determines very precisely the field of application of the models produced. Before using any of them it is important to have a clear idea of the biological question for which the model should provide an answer and to verify whether the selected method is sufficiently accurate for the problem at hand. The latter depends upon several factors, and the aim of this book is to provide students and experimental researchers with a guide through the available methods, their underlying assumptions, their limitations, and their expected accuracy in different applications.

Before entering into the heart of the problem, we will discuss the relationship between sequence, structure, and function in proteins, to familiarize the reader with the features of a protein structure (Chapter 1) and with methods and initiatives aimed at evaluating the effectiveness and limitations of prediction methods, and their expected accuracy (Chapter 2).

In Chapter 3 we discuss the extent to which physics can help us compute the structure of a protein on the basis of the knowledge of its chemical structure. The possibility of predicting the structure of a protein from scratch using this approach is, unfortunately, out of our reach at present, and therefore we will explore alternative, knowledge based methods, in subsequent chapters. In Chapter 4 we will survey how a three-dimensional model of a protein can be built on the basis of its evolutionary relationship with a protein of known structure and how to detect the existence of such a relationship. Next, we will discuss how models of proteins can be built taking advantage of the observation that apparently unrelated proteins can share a similar overall structure (Chapter 5) or that, in any case, they share local structure similarities (Chapter 6). Methods which can be used to infer some structural properties of proteins, even if they do not provide us with their complete three-dimensional structure will be surveyed in Chapter 7. These methods are very useful for guiding the construction and assessing the reliability of atomic models obtained with other techniques.

Special treatment must be reserved to membrane proteins. Their peculiar properties, dictated by the environment in which they reside, enable different

methods and strategies to be used for their prediction; these will be discussed in Chapter 8.

Finally, a description of a few successful examples of protein structure modeling will be discussed in the last chapter. The list could be much longer, the literature abounds with scientific reports in which structure prediction is used to guide experiments and to interpret data in the light of a model of the proteins of interest. The examples chosen here partially reflect the personal preference of the author and partially are meant to touch different aspects of the field of protein structure prediction.

The aim of the book is to convince the reader that protein modeling can be extremely useful, if it is performed carefully and with full understanding of the techniques, and that it should be application driven. There is no point in constructing a very approximate model of a protein and later to use it to derive detailed properties about catalytic mechanisms or interaction details; similarly, it is not wise to employ very sophisticated techniques to obtain a model of a protein that, for technical reasons, cannot be used for guiding experiments or for casting light on important biological questions.

Rome, November 2005 *Anna Tramontano*

Acknowledgments

I would like to express my profound gratitude to all my friends and colleagues who offered advice during the preparation of this book. I am deeply grateful to Domenico Cozzetto who took the time to read early drafts of this book and offer many valuable comments and criticisms.

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

Introduction

Proteins are the functioning molecules of living organisms; they play key roles in most physiological processes, for example metabolism, transport, immune response, signal transduction, and cell cycle. These linear polymers of amino acids perform this impressive variety of functions by assuming a well defined three-dimensional structure, which enables them to accomplish highly specific molecular functions. In these complex structures, amino acids that are far apart in the linear sequence can come close together in space and participate in the formation of highly specialized catalytic sites, ligand-binding pockets, or interfaces able to recognize, bind and transmit information to other macromolecules.

Genomic projects are providing us with the linear amino acid sequence of hundreds of thousands of proteins. If only we could learn how each and every one of these folds in three-dimensions we would have the complete part list of an organism and could face the challenge of understanding how these parts assemble in a cell. This is not only an intellectual challenge, it has enormous practical implications. Malfunctioning of proteins is the most common cause of endogenous diseases and the action of pathogens is usually mediated by their proteins. Most life-saving drugs act by interfering with the action of a faulty or foreign protein by keeping it from performing its function. Sometimes the drug competes directly with the substrate occupying the site where action occurs, sometimes it binds in a different location of the protein surface, causing a structural effect that modifies the geometry of the active site and makes the protein unable to perform its biochemical function. So far, most drugs have been discovered by trial-and-error, testing thousands of randomly selected compounds for their ability to interfere with the function of a protein or with a biological process. The active compounds are later modified to try and endow them with desirable properties such as higher specific activity, lack of toxicity, ability to reach the correct cellular compartment, favorable metabolic properties, ease of chemical synthesis, etc. The rate of success of this process is rather low, it is estimated that only one out of a thousand active molecules makes it to the pharmacy bench.

Our lack of a complete understanding of the complex interplay between different proteins has several implications for our welfare. The drugs we use might not be aimed at the best target, i.e. other proteins participating in the same biological process might be better targets. Drugs usually have undesirable side-effects that can only be assessed in clinical trials, because we cannot evaluate beforehand

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

whether they bind and interfere with molecules involved in different processes. Differences between the genetic background of individuals might cause the protein of a specific individual to be less sensitive to the effect of a drug or cause the drug to have more serious side effects. All this makes the drug-development process very time consuming and far from optimum. In an ideal situation, if we had a deep structural understanding of our complete parts list, we could screen drugs in a virtual in-silico system and detect a substantial proportion of these problems beforehand, increasing substantially the effectiveness of medical approaches. To achieve this goal we need to know the structure of the proteins involved at a very accurate level of detail.

Experimental methods can provide us with knowledge of the precise arrangement of every atom of a protein. The most effective of these are X-ray crystallography and NMR spectroscopy. X-ray crystallography, however, requires the protein or the protein complex under study to form a reasonably well ordered crystal, a feature that is not universally shared by proteins. NMR spectroscopy needs proteins to be soluble and there is a limit to the size of protein that can be studied. The structural information on some proteins can be very difficult to obtain experimentally. For example, proteins embedded in hydrophobic biological membranes are both difficult to crystallize and insoluble in polar solvents. Yet, these proteins play fundamental roles in biology because they are involved in the process of communication between cells and in the uptake of external molecules. Both X-ray crystallography and NMR are time-consuming techniques and we cannot hope to use them to solve the structures of all the proteins of the universe in the near future. As will be discussed, we know that the structure of a protein depends solely on its amino acid sequence, therefore we would like to develop theoretical approaches to infer (or predict, as we say) the structure of a protein from its amino acid sequence.

For decades, the problem of deciphering the code that relates the amino acid sequence of a protein and its native three-dimensional structure has been the subject of innumerable investigations and, despite the many frustrations caused by its elusiveness, interest in the problem is not fading away. On the contrary. What stands in our way, notwithstanding all these efforts, is the complexity of protein structures. In a three-dimensional protein structure, thousands of atoms are held together by weak forces and give rise to a conformation which is only marginally stable. The consequence of this, as we will discuss, is that it is very unlikely that we can use our understanding of the laws of physics to compute the native functional structure of a protein in the foreseeable future. We have at our disposal, however, the experimentally solved structures of a reasonable number of proteins, currently a few thousand. They represent solved instances of our problem and we can hope to extract heuristic rules from their analysis. This is the topic of this book – it describes what we can learn from the analysis of known protein structures and how this knowledge can be used to attempt the prediction of unknown protein structures. As we will see, this approach has led to several methods for protein structure prediction, some reaching a respectable level of accuracy and reliability. Usually, however, the accuracy of a model is not comparable with that achievable

by experimental methods, and it is therefore important to understand the limitations of the methods.

Now that the sequences of many proteins are available to us and computers are becoming increasingly powerful, it is possible to run high throughput modeling experiments and predict the structures of thousands of proteins automatically. In other, more difficult instances, models must be built manually, taking into account other data, especially experimental, and each step of the process must be accurately analyzed for the possible presence of errors. In both instances, but even more in the latter, it is important to ask a few questions before approaching the problem of predicting the structure of a protein:

- What will be the application of the model?
- Is the expected accuracy of the obtainable model appropriate for the task?
- Can we devise experiments to verify the correctness of our model before we use it to infer the properties of the target protein?

Models of proteins can, undoubtedly, be very useful for several applications and we will describe some examples in which they have been instrumental in achieving better understanding of a difficult biological problem, but it should always be kept in mind that there are limits to their accuracy. Admittedly, experimental determinations of protein structures are also affected by errors, as are all experimental data, but at present, except for a handful of cases, the accuracy of an experimental structure is far superior to that achievable with modeling methods.

The situation is, therefore, that we have a limited number of protein structures determined with high accuracy and less detailed information derived by modeling on a much larger number of proteins. There is no doubt that the most efficient strategy is for the two fields to complement each other in exploring the protein structure space. We should optimize efforts and solve experimentally the structure of proteins selected in a way that increases the chances that modeling methods can use them to produce reasonably accurate models of many others. Worldwide initiatives, denoted structural genomics projects, are indeed following this idea and are producing the structures of a large number of proteins selected with the aim of providing representative examples of the protein structural space. The task of protein structure modeling is to make the best use of the available experimental protein structure information to infer the structures of as many of the remaining proteins as possible.

Traditionally, protein structure prediction methods have been categorized according to the nature of the relationship between the protein to be modeled and the available proteins of known structure. Comparative or homology modeling is based on the observation that evolutionarily related proteins preserve their structural features during evolution and is the method of choice when a clear evolutionary relationship can be detected between the target protein and one or more proteins of known structure. Even proteins with no clear evolutionary relationship often share a similar structure and methods devoted to detecting these and exploiting this information to model a protein are called "fold-recognition methods". Finally, in all other instances, the observation that some structural features are shared between

proteins not falling into any of the two previous categories, gave rise to methods known as "fragment-based methods". The boundaries between the three methodologies are, as we will see, becoming less defined, and methods are cross-fertilizing each other in several ways. Nevertheless, here we will still follow the traditional subdivision between methods, because we believe that it makes it easier to navigate through the many aspects of this fascinating problem.

1
Sequence, Function, and Structure Relationships

1.1
Introduction

Life is the ability to metabolize nutrients, respond to external stimuli, grow, reproduce, and, most importantly, evolve. Most of these functions are performed by proteins, organic macromolecules involved in nearly every aspect of the biochemistry and physiology of living organisms. They can serve as structural material, catalysts, adaptors, hormones, transporters, regulators. Chemically, proteins are linear polymers of amino acids, a class of organic compounds in which a carbon atom (called Cα) is bound to an amino group ($-NH_2$), a carboxyl group (–COOH), a hydrogen atom (H), and an organic side group (called R). The physical and chemical properties unique to each amino acid result from the properties of the R group (Figure 1.1).

In a protein, the amino group of one amino acid is linked to the carboxyl group of its neighbor, forming a peptide (C–N) bond. There are two resonance forms of the peptide bond (i.e. two forms that differ only in the placement of electrons), as illustrated in Figure 1.2. Atoms involved in single bonds share one pair of electrons whereas two pairs are shared in double bonds. The latter are planar, i.e. not free to rotate. The resonance between the two forms shown in Figure 1.2 makes the peptide bond intermediate between a single and a double bond and, as a consequence, all peptide bonds in protein structures are found to be almost planar – the Cα, C, and O atoms of one amino acid and the N, H, and Cα of the next lie on the same plane. Although this rigidity of the peptide bond reduces the degrees of freedom of the polypeptide, the dihedral angles around the N–Cα and the Cα–C bonds are free to vary and their values determine the conformation of the amino acid chain.

Proteins assume a three-dimensional shape which is usually responsible for their function. The consequence of this tight link between structure and function and of the evolutionary pressure to preserve function has a very important effect – in contrast with ordinary polymers (e.g. polypeptides with random sequences) that typically form amorphous globules, proteins usually fold to unique structures. In other words they spontaneously assume a unique three-dimensional structure specified, as we will see, by their amino acid sequence. For example, enzymes accelerate chemical reactions by stabilizing their high energy intermediate and this

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

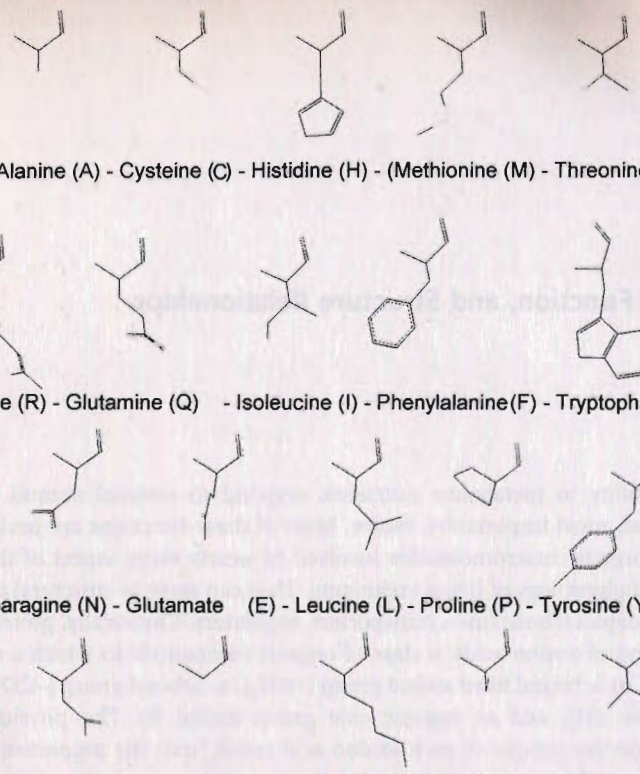


Figure 1.1 The twenty naturally occurring amino acids.

is achieved by correct relative positioning of appropriate chemical groups. Our body contains many proteins that catalyze the hydrolysis of peptide bonds in proteins (the inverse of the polymerizing reaction used to build proteins) to provide the body with a steady supply of amino acids. The substrates of these reactions are either proteins from the diet or "used" proteins inside the body. Digestion begins in the stomach where the acidic environment unfolds, i.e. destructures, the proteins and an enzyme called pepsin (Figure 1.3) starts chopping the proteins into pieces. Later, in the intestines, several protein-cutting enzymes, for example trypsin (also shown in Figure 1.3), cut the protein chains into shorter pieces. In subsequent steps, other enzymes reduce these shorter pieces to single amino acids.

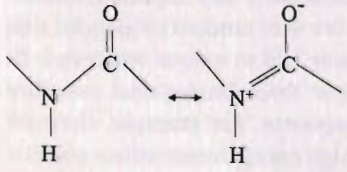
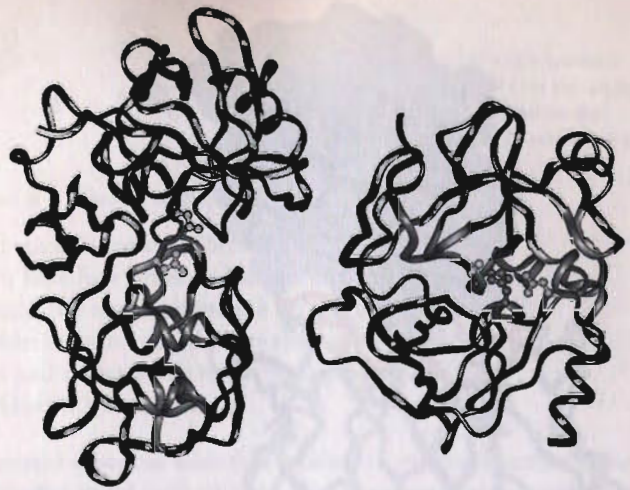


Figure 1.2 The two resonance structures of the peptide bond. Because of delocalization of the electrons, the C–N bond has the character of a partial double bond and this limits its freedom of rotation.



IGDEPLENYL DTEYFGTIGI GTPAQDFTVI FDTGSSNLWV
PSVYCSSLAC SDHNQFNPDD SSTFEATSQE LSITYGTGSM
TGILGYDTVQ VGGISDTNQI FGLSETEPGS FLYYAPFDGI
LGLAYPSISA SGATFVFDNL WDQGLVSQDL FSVYLSSNDD
SGSVVLLGGI DSSYYTGSLN WVPVSVEGYW QITLDSITMD
GETIACSGGC QAIVDTGTSL LTGPTSAIAN IQSDIGASEN
SDGEMVISCS SIASLPDIVF TINGVQYPLS PSAYILQDDD
SCTSGFEGMD VPTSSGELWI LGDVFIRQYY TVFDRANNKV
GLAPVA

IVGGYTCGAN TVPYQVSLNS GYHFCGGSLI NSQWVVSAA
CYKSGIQVRL GEDNINVVEG NEQFISASKS IVHPSYNSN
LNNDIMLIKL KSAASLNSRV ASISLPTSCA SAGTQCLIS
WGNTKSSGTS YPDVLKCLKA PILSDSSCKS AYPGQITSN
FCAGYLEGGK DSCQGDSGGP VVCSGKLQGI VSWGSGCAQ
NKPGVYTKVC NYVSWIKQTI ASN

Figure 1.3 The three-dimensional structures of pepsin (left, PDB code: 1PSN) and trypsin (right, PDB code: 3TPI). These two enzymes cleave peptide bonds with a different mechanism. The first uses two aspartic acids, the second a triad formed by a histidine, a serine, and an aspartic acid. Their amino acid sequence is shown at the bottom of the figure in a one-letter code. Note that, for both enzymes, the amino acids forming the active site (underlined) are distant in the linear sequence and are brought together by the three-dimensional structure of the enzymes.

Both pepsin and trypsin belong to a class of enzymes called proteases, the first performs its job by taking advantage of the presence, in a cleft of the protein structure of two residues of aspartic acid; in the second catalysis is achieved by cooperation of three amino acids, a serine, a histidine, and an aspartic acid. Amino acids near the active site are responsible for recognition and correct positioning of the substrate. These functional amino acids, far apart in the linear amino acid sequence (Figure 1.3), are brought together in exactly the right position by the protein three-dimensional structure.

Similarly, recognition of foreign molecules is mediated by several proteins of the immune system, the most popular being antibodies. Antibodies bind other molecules, called antigens, by means of an exposed molecular surface complementary to the surface of the antigen, which can be a protein, a nucleic acid, a polysaccharide, etc. The binding surface is formed by amino acids from different parts of the molecule (Figure 1.4).

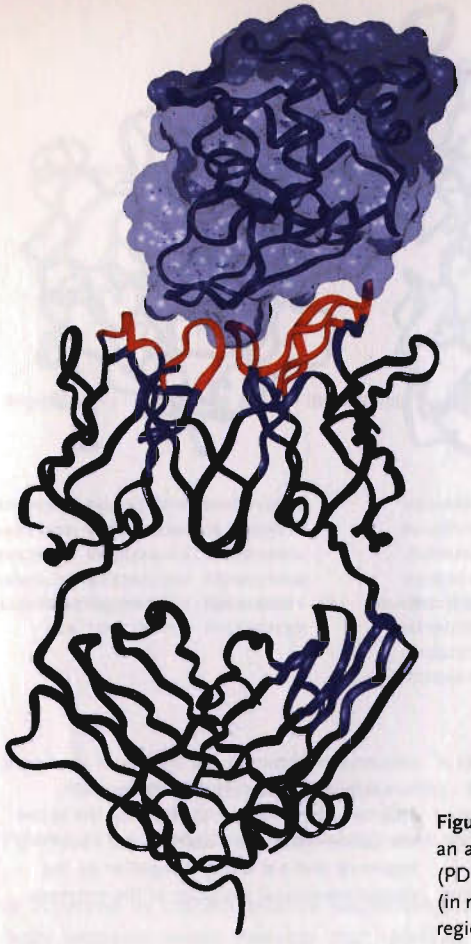


Figure 1.4 The three-dimensional structure of an antibody bound to its cognate molecule (PDB code: 3HFL). Note that the binding region (in red) is formed by amino acids from different regions of the linear sequence.

1.2 Protein Structure

Most readers will already be familiar with the basic concepts of protein structure; we will, nevertheless, review here some important aspects of this subject. The sequence of amino acids, i.e. of the R-groups, along the chain is called the primary structure. Secondary structure refers to local folding of the polypeptide chain. Tertiary structure is the arrangement of secondary structure elements in three dimensions and quaternary structure describes the arrangement of a protein's subunits. As we have already mentioned, the peptide bond is planar and the dihedral angle it defines is almost always 180°. Occasionally the peptide bond can be in the cis conformation, i.e. very close to 0°.

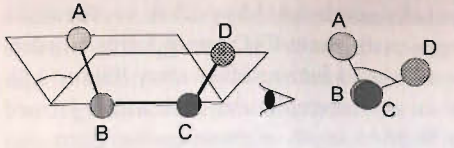


Figure 1.5 A dihedral angle between four points A, B, C, and D is the angle between two planes defined by the points A, B, C and B, C, D, respectively.

Question: What is a dihedral angle?

»The dihedral angle is the angle between two planes. In practice, if you have four connected atoms and you want to measure the dihedral angle around the central bond, you orient the system in such a way that the two central atoms are superimposed and measure the resulting angle between the first and last atom (Figure 1.5).«

The simplest arrangement of amino acids that results in a regular structure is the alpha helix, a right-handed spiral conformation. The structure repeats itself every 5.4 Å along the helix axis. Alpha helices have 3.6 amino acid residues per turn and

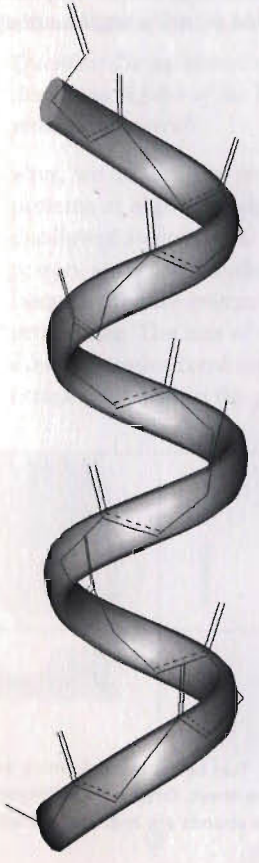


Figure 1.6 An alpha helix. Each backbone oxygen atom is hydrogen-bonded to the nitrogen of a residue four positions down the chain.

the separation of the residues along the helix axis is 5.4/3.6 or 1.5 Å, i.e. the alpha helix has a rise per residue of 1.5 Å. Every main-chain C=O group forms a hydrogen bond with the NH group of the peptide bond four residues away (Figure 1.6).

Let us recall that a hydrogen bond is an intermolecular interaction formed between a hydrogen atom covalently bonded to an electronegative atom (for example oxygen or nitrogen) and a second electronegative atom that serves as the hydrogen-bond acceptor. The donor atom, the hydrogen, and the acceptor atom are usually co-linear. The alpha helix has 3.6 residues per turn and thirteen atoms enclosed in the ring formed by the hydrogen bond, it can also be called a 3.6(13) helix. Another type of helix is observed in protein structures, although much more rarely; this is the 3(10) helix. This arrangement contains three residues per turn and ten atoms in the ring formed by the hydrogen bond. In alpha helices, the peptide planes are approximately parallel with the helix axis, all C=O groups point in one direction, and all N–H groups in the opposite direction. Because of the partial charge on these groups, negative for CO and positive for NH, there is a resulting dipole moment in the helix. Side-chains point outward and pack against each other. The dipoles of a 3(10) helix are less well aligned and the side-chain packing less favorable, therefore it is usually less stable. Typically, in alpha helices the angles around the N–Cα and Cα–C bonds, called φ and ψ angles, are approximately –60° and –50°, respectively.

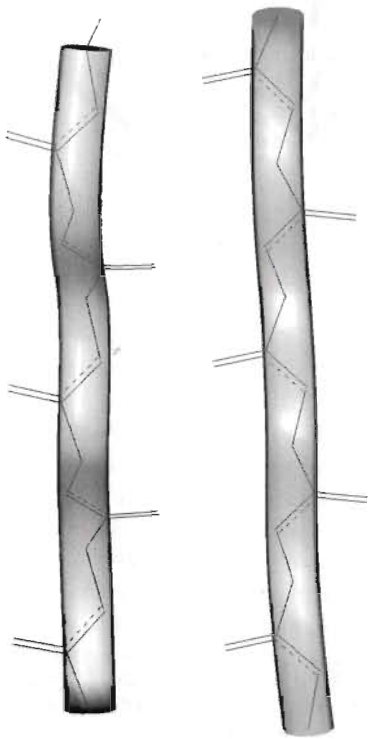


Figure 1.7 Two beta strands forming an antiparallel beta sheet. Oxygen and nitrogen atoms of different strands are hydrogen bonded to each other.

Another secondary structure element commonly observed in proteins is the beta sheet, an arrangement of two or more polypeptide chains (beta strands) linked in a regular manner by hydrogen bonds between the main chain C=O and N–H groups. The R groups of neighboring residues in a beta strand point in opposite directions forming a layered structure (Figure 1.7). The strands linked by the hydrogen bonds in a beta sheet can all run in the same direction (parallel sheet) or in opposite directions (antiparallel sheet). Beta sheets can be mixed, including both parallel and antiparallel pairs of strands. Most beta sheets found in proteins are twisted – each residue rotates by approximately 30° in a right-handed sense relative to the previous one.

The plot shown in Figure 1.8 is called a Ramachandran plot. This can be obtained by considering atoms as hard spheres and recording which pairs of φ and ψ angles do not cause the atoms of a dipeptide to collide. Allowed pairs of values are represented by dark regions in the plot whereas sterically disallowed regions are left white. The lighter areas are obtained by using slightly smaller radii of the spheres, i.e. by allowing atoms to come a bit closer together. Disallowed regions usually involve steric hindrance between the first carbon atom of the side-chain, the Cβ, and main-chain atoms. As we will see, the amino acid glycine has no side-chain and can adopt φ and ψ values that are unfavorable for other amino acids.

Question: Do we observe amino acids with dihedral angles in disallowed regions of the Ramachandran plot in experimental protein structures?

»Yes, we do. Even in very well refined crystal structures of proteins at high resolution, some φ and ψ angles fall into disallowed regions. The reader should keep in mind that the reason some combinations of angles are rarely observed is because they are energetically disfavored, not mathematically impossible. The loss of energy because of an unfavorable dihedral angles combination can be compensated by other interactions within the protein.«

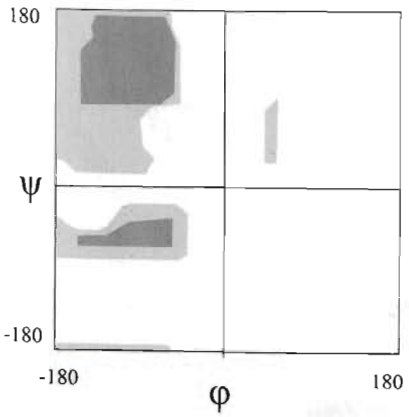


Figure 1.8 A Ramachandran plot is a graph reporting the values of phi and psi angles in protein structures. Darker areas indicate favorable combinations of angles, lighter gray areas are less favored, but still possible.

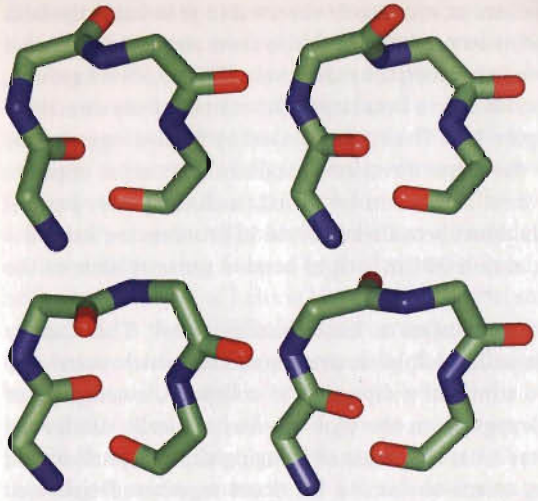


Figure 1.9 The four types of beta turn described in Table 1.1, types I and I′ are shown on the top, types II and II′ on the bottom.

Regions without repetitive structure connecting secondary structure elements in a protein structure are called loops. The amino acid chain can reverse its direction by forming a reverse turn characterized by a hydrogen bond between one main chain carbonyl oxygen and the N–H group 3 residues along the chain (Figure 1.9). When such a secondary structure element occurs between two anti-parallel adjacent beta strands in a beta sheet is called a beta hairpin. Reverse turns are classified on the basis of the ϕ and ψ angles of the two residues in their central positions as shown in Table 1.1. Note that some turns require that one of their amino acids has ϕ and ψ angles falling in disfavored regions of the Ramachandran plot.

Table 1.1 Turns are regions of the protein chain that enable the chain to invert its direction. The ϕ and ψ angles of some commonly occurring turns are listed.

Turn type	ϕ_1	ψ_1	ϕ_2	ψ_2
I	–60	–30	–90	0
I′	60	30	90	0
II	–60	120	80	0
II′	60	–120	–80	0

Proteins can be formed from only alpha helical or from only beta sheet elements, or from both; the association of these elements within a single protein chain is called tertiary structure. Certain arrangements of two or three consecutive secondary structures (alpha helices or beta strands), are present in many different protein

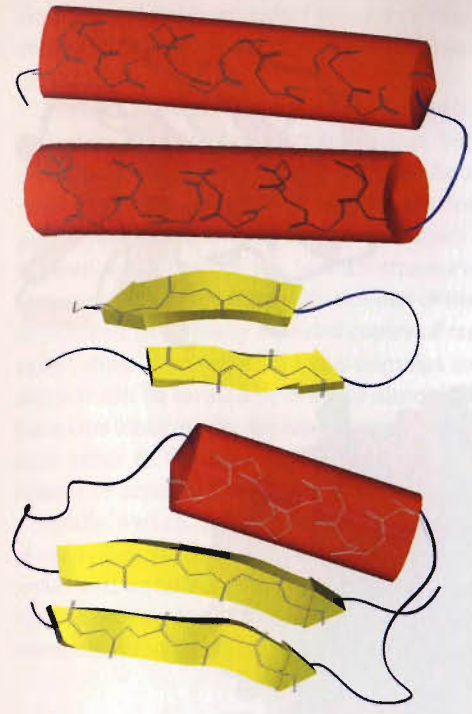


Figure 1.10 Supersecondary structures: alpha–loop–alpha, beta hairpin, and beta-alpha-beta unit.

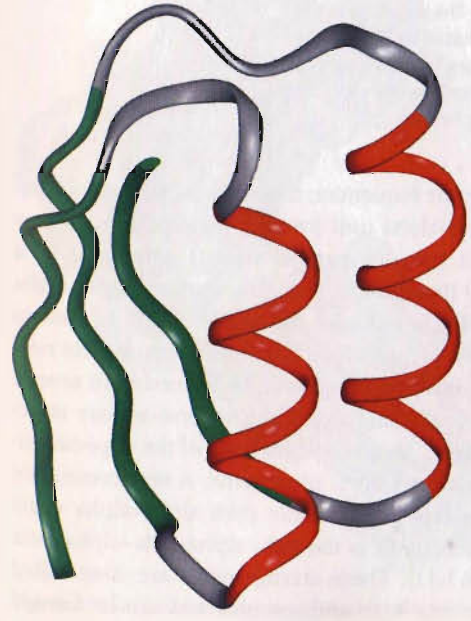


Figure 1.11 The Rossmann fold. The figure shows a region of the succinyl-Coa synthetase enzyme from the bacterium *Escherichia coli* (PDB code: 2SCU).

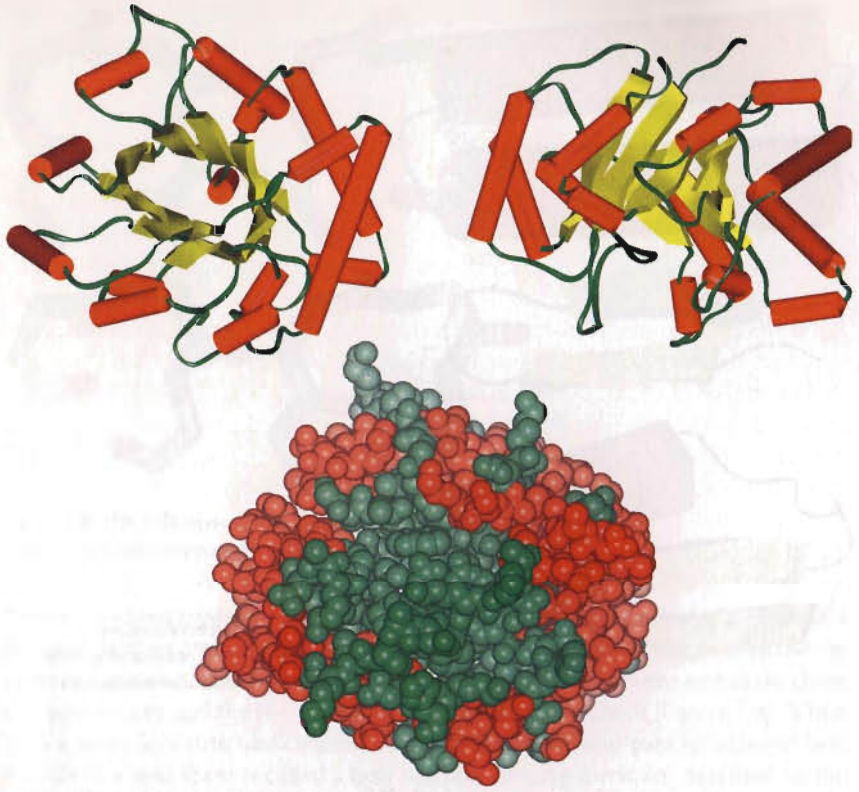


Figure 1.12 A TIM barrel (PDB code: 8TIM). The structure at the top left is the same as that at the top right rotated by 90° around an horizontal axis. On the bottom the structure is shown with all its non-hydrogen atoms. Atoms in green belong to the central beta barrel, atoms in red to the surrounding helices.

structures, even with completely different sequences; these are called supersecondary structures. They include the alpha–alpha unit (two antiparallel alpha helices joined by a turn); the beta–beta unit (two antiparallel strands connected by a hairpin); and the beta–alpha–beta unit (two parallel strands, separated by an alpha helix antiparallel to them (Figure 1.10). Sometimes the term "motif" is used to describe these supersecondary structures. Supersecondary structures are not necessarily present in a protein structure, however, which can be formed from several alpha helices or beta strands without containing any of the supersecondary structures described above. On the other hand, some combinations of the supersecondary structural motifs are observed relatively often in proteins. A very commonly found arrangement of helices is the four-helix bundle (two alpha–alpha units connected by a loop). Another common motif is the beta–alpha–beta–alpha–beta unit, called the Rossman fold (Figure 1.11). These arrangements are often called domains or folds. Some folds can be very large and complex and can be formed

from several supersecondary structures. One example is the TIM barrel fold; this is shared by many enzymes and formed from several beta–alpha–beta units (Figure 1.12).

Another layer of organization of protein structure is the domain level. The definition of a domain is rather vague. Some confusion also arises because the term is often also used in the context of the amino acid sequence, rather than of its three-dimensional structure. In general a domain can be defined as a portion of the polypeptide chain that folds into a compact semi-independent unit. Domains can be seen as "lobes" of the protein structure that seem to have more interaction between themselves than with the rest of the chain (Figure 1.13). Several proteins are formed from many repeated copies of one or a few domains; such proteins are called mosaic proteins and the domains are often referred to as "modules". A domain can be formed by only (or almost only) alpha helices or beta sheets, or by their combination. In the latter case the helices and strands can be packed against each other in the beta–alpha–beta supersecondary arrangement (alpha/beta domains) or separated in the structure (alpha + beta domain).

Finally, we talk about architecture of a protein when we consider the orientations of secondary structures and their packing pattern, irrespective of their sequential order, and we talk of protein topology when we also take into account the nature of the connecting loops and, therefore, the order in which the secondary structure elements occur in the amino acid sequence.

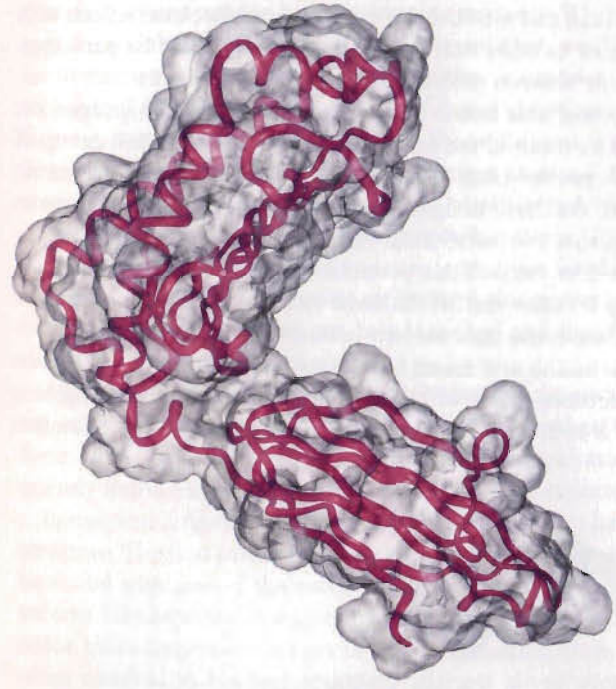


Figure 1.13 A two-domain protein chain (PDB code: 1HSA).

1.3 The Properties of Amino Acids

There are twenty naturally occurring amino acids. They can play different roles and it is important to survey their properties to be able to analyze and ultimately attempt to predict the structure and function of a protein.

The smallest amino acid is glycine, the side-chain of which is just a hydrogen atom. The lack of a side-chain makes this amino acid very flexible. We have already mentioned that this amino acid can assume "unusual" ϕ and ψ angles. We also saw that the structural requirements of turns often need an amino acid in this conformation and indeed these positions are often occupied by glycines. The observation that a glycine is always present in a given position in a family of evolutionarily related proteins often points to the presence of a tight turn in the region. The flexibility of glycine also implies that the loss of entropy associated with restricting its conformation in a protein structure is higher than for other amino acids, and the absence of a side-chain makes it less likely for this amino acid to establish favorable interactions with surrounding amino acids. Glycines are, therefore, rarely observed in both alpha helices or beta sheets.

The next amino acid, in order of size, is alanine. Here the side-chain is a CH_3 group. It is a small hydrophobic amino acid, without any reactive group, and rarely involved in catalytic function. Its small non polar surface and its hydrophobic character suggest, however, that this amino acid can be exposed to solvent, without large loss of entropy, and can also establish favorable hydrophobic interactions with other hydrophobic surfaces. In other words, it is an ideal amino acid for participating in interacting surfaces between proteins that associate transiently.

Cysteine is a small hydrophobic amino acid that can form disulfide bridges, i.e. covalent bonds arising as a result of the oxidation of the sulfhydryl (SH) group of the side-chains of two cysteine units when they are in the correct geometric orientation (Figure 1.14). Disulfide bridges enable different parts of the chain to be covalently bound. Because the intracellular environment is reducing, disulfide bridges are only observed in extracellular proteins. Cysteine can also coordinate metals and its SH group is rather reactive. In some viral proteases it takes the role of serine in serine protease active sites we have already described.

Serine is a small polar amino acid found both in the interior of proteins and on their surfaces. It is sometimes found within tight turns, because of its small size and its ability to form a hydrogen bond with the protein backbone. It is often

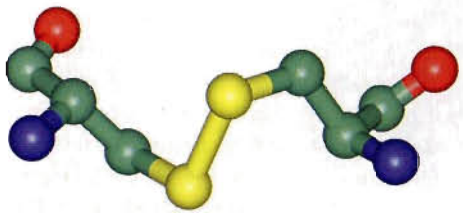


Figure 1.14 A disulfide bond. The yellow atoms are sulfur atoms.

observed in active sites, where it can act as a nucleophile as already mentioned for serine proteases. Another important property of this amino acid is that it is a substrate for phosphorylation – enzymes called protein kinases can attach a phosphate group to its side-chain. This plays important roles in many cellular processes and in signal transduction.

Another relatively small amino acid, rather similar to serine, is threonine. This amino acid can also be part of active sites and can be phosphorylated. An important difference, though, is that threonine is "beta branched", i.e. it has a substituent on its beta carbon and this makes it less flexible and less easy to accommodate in alpha helices. Beta-branched amino acids are indeed more often found in beta sheets.

Asparagine and glutamine are polar amino acids that generally occur on the surface of proteins, exposed to an aqueous environment, and frequently involved in active sites. Asparagine, for example, is found as a replacement for aspartate in some serine proteases. One peculiar property of asparagine is that it is often found in the left-handed conformation (positive ϕ and ψ angles) and can therefore play a role similar to that of glycine in turns. This is possibly because of its ability to form hydrogen bonds with the backbone.

Proline is unique because it is an imino acid rather than an amino acid. This simply means that its side-chain is connected to the protein backbone twice, forming a five-membered nitrogen-containing ring. This restricts its conformational flexibility and makes it unable to form one of the two main-chain hydrogen bonds that other amino acids can form in secondary structure elements; it is, therefore, often found in turns in protein structures. When it is in an alpha helix, it induces a kink in the axis of the helix. It is not a very reactive amino acid, but plays an important role in molecular recognition – peptides containing prolines are recognized by modules that are part of many signaling cascades. Proline can be found in the cis conformation (i.e. with the angle around the peptide bond close to 0° rather than 180°). The main chain nitrogen atoms of the other amino acids are bound to a hydrogen and a carbon atom whereas the situation in proline is more symmetrical with the atom bound to two carbon atoms. This means that the energy difference between the *cis* and *trans* conformations is smaller for this amino acid.

Leucine, valine, and isoleucine are hydrophobic amino acids, very rarely involved in active sites. The last two are beta branched and therefore often found in beta sheets and rarely in alpha *helices*.

Aspartate and glutamate are negatively charged amino acids, generally found on the surface of proteins. When buried, they are involved in salt bridges, i.e. they form strong hydrogen bonds with positively charged amino acids. They are frequently found in protein active sites and can bind cations such as zinc.

Lysine and Arginine are positively charged and can have an important role in structure. The first part of their side-chain is hydrophobic, so these amino acids can be found with part of the side-chain buried, and the charged portion exposed to solvent. Like aspartate and glutamate, lysine and arginine can form salt bridges and occur quite frequently in protein active or binding sites. They are, furthermore, often involved in binding negatively charged phosphates and in the interacting surfaces of DNA- or RNA-binding proteins.

At physiological pH, histidine can act as both a base or an acid, i.e. it can both donate and accept protons. This is an important property that makes it an ideal residue for protein functional centers such as the serine protease catalytic triad. Histidine can, furthermore, bind metals (e.g. zinc). This property is often exploited to simplify purification of proteins cloned and expressed in heterologous systems. The addition of a tail of histidines to the protein of interest confers on the protein the ability to chelate metals and this engineered property can be exploited for purifying the protein.

Methionine has a long and flexible hydrophobic side-chain. It is usually found in the interior of proteins. Like cysteine, it contains a sulfur atom, but in methionine the sulfur atom is bound to a methyl group, which makes it much less reactive.

Phenylalanine, tryptophan, and tyrosine are aromatic amino acids. The term "aromatic" was used by chemists to describe molecules with peculiar odors long before their chemical properties were understood. In chemistry, a molecule is called aromatic if it has a planar ring with $4n + 2$ π-electrons where n is a non-negative integer (Hückel's Rule). In practice, these molecules, the prototype of which is benzene, have a continuous orbital overlap that gives them special optical properties. For example, tryptophan absorbs light at 280 nm and this property is routinely used to measure the concentration of proteins in a solution (assuming the protein contains at least one tryptophan). Also, if an aromatic residue is held rigidly in space in an asymmetric environment, it absorbs left-handed and right-handed polarized light differently. This effect, which can be measured by circular dichroism spectroscopy, is therefore sensitive to overall three-dimensional structure and can be used to monitor the conformational state of a protein. Another important property of amino acids with aromatic side-chains is that they can interact favorably with each other. The face of an aromatic molecule is electron-rich while the hydrogen atoms around the edge are electron-poor. This implies that off-set face-to-face and edge-to-face interactions between aromatic rings have both hydrophobic and electrostatic components. Tyrosine is also a substrate for phosphorylation, similarly to serine and threonine, although the enzymes responsible for phosphorylation of tyrosine are different from those that phosphorylate serine and threonine.

1.4 Experimental Determination of Protein Structures

Two experimental techniques are used to determine the three-dimensional structure of macromolecules at the atomic level – X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Although it is beyond the scope of this book to describe the details of these techniques, which are rather complex both theoretically and experimentally, it is important to have some basic understanding of their results, because, as we will see, most methods for prediction of protein structure are based on existing structural data.

X-ray crystallography is based on the fact that an ordered ensemble of molecules arranged in a crystal lattice diffracts X-rays (the wavelengths of which are of the order of interatomic distances) when hit by an incident beam. The X-rays are dispersed by the electrons in the molecule and interfere with each other giving rise to a pattern of maxima and minima of diffracted intensities which depends upon the position of the electrons (and hence of the atoms) in the ordered molecules in the crystal. The electron density of the protein, i.e. the positions of the protein atoms, determines the diffraction pattern of the crystal, that is the magnitudes and phases of the X-ray diffraction waves, and vice versa, through a Fourier transform function. In practice:

$$\varrho(x,y,z) = \frac{1}{V}\sum_{hkl}\vec{F}_{hkl} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}F(h,k,l)e^{-2\pi i(hx+ky+lz)}$$

where $\varrho(x,y,z)$ is the electron density at position (x,y,z), V is the volume, $\vec{F}(h,k,l)$ is the vector describing the diffracted waves in terms of their amplitudes $F(h,k,l)$ and phases (the exponential complex term). The electron density at each point depends on the sum of all of the amplitudes and phases of each reflection. If we could measure the amplitude and phase of the diffracted waves, we could relatively easily compute the exact relative location of each atom in the diffracting molecules. Unfortunately the phase of the diffracted waves cannot be measured and, therefore, we must use "tricks" to guess their approximate value and reconstruct the image of the diffracting molecule.

In effect, three methods are used to estimate the phases. Direct methods consist in using all possible values for the phases in the Fourier transform equation until an interpretable electron density is found; this is feasible for small molecules only. Interference-based methods can make use of multiple isomorphous replacement or anomalous scattering techniques. The first derives the phase by comparing the diffraction pattern of a protein crystal with that of crystals identical to the original one but for the presence of "heavy" atoms (i.e. atoms with many electrons and, therefore, very strong diffracting power) in specific positions of the molecules. The "anomalous scattering" technique instead derives initial phases by measuring diffraction data at several different wavelengths near the absorption edge of a heavy-atom. Finally, if we have a reasonable model for the molecule in the crystal, we can resort to the "molecular replacement" technique which computes approximated phases for the molecule in the crystal on the basis of the position of the atoms in the model. The availability of a high-quality three-dimensional model for a protein can therefore also be instrumental in obtaining its experimentally determined structure.

Given the diffracted intensities of a protein crystal and a set of "good" estimated phases, we can calculate the electron density that formed the observed pattern and position the atoms of the protein in the computed electron density (Figure 1.15). Important aspects of the whole procedure are that the protein under examination forms well ordered, well diffracting crystals and that the phase estimation procedure is successful in generating an interpretable electron-density map.

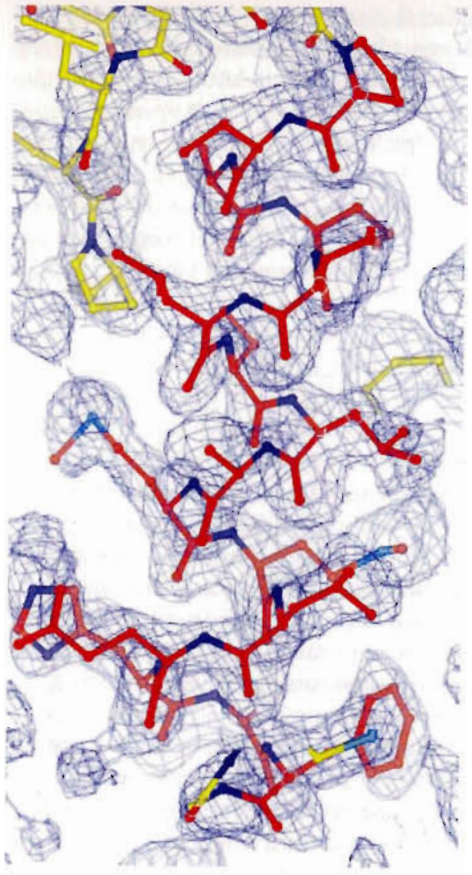


Figure 1.15 An electron-density map derived from an X-ray crystallography experiment. The atoms can be positioned in the map as shown in the figure, revealing an alpha helical structure.

Question: Is the quality of an X-ray determination of a protein structure comparable to that for small organic molecules?

»The quality of the structural data that can be obtained by protein crystallography is nowhere near the accuracy with which crystal structures of small molecules can be determined. This is because proteins can assume many different, although closely related, conformations and this limits the order of the molecules within the crystal. Also, protein crystals are usually only about half protein – the other half is occupied by solvent molecules. As we will see, the accuracy of small molecule crystallography can be used to derive parameters useful in modeling procedures.«

Just as in every experiment, in protein crystallography also the quality of the results improves with the ratio of the amount of data collected (the diffraction intensities) and the number of properties estimated (the positions of the atoms). In crystallography, the inverse of this ratio is expressed by the term "resolution" which is expressed in Angstroms (1 Angstrom = 10^{-10} m). The lower is the resolution the better is the quality of the structure. A resolution of approximately 3Å enables secondary structural elements and the direction of the polypeptide chain to be clearly identified in the electron density map; with a resolution of 2.5Å the side-chains can be built into the map with reasonable precision.

Hydrogen atoms do not diffract very well, because they only have one electron, and they are usually not detectable by X-ray crystallography unless the resolution is really very good, approximately 1.0Å. This implies that, given a crystal structure with good but not exceptional resolution, we can only deduce the presence of hydrogen bonds by the position of the donor and acceptor atoms.

After reconstructing the structure, we can back compute the expected diffraction pattern and compare it with that observed. The *R* factor indicates how much the two patterns (theoretical and experimental) differ and is expressed as a percentage. This factor is linked to the resolution. As a rule of thumb, a good structure should have an *R* factor lower than the resolution divided by 10 (i.e. ≤30% for a 3.0Å resolution structure, ≤20% for a 2.0Å structure, etc). To avoid any bias, it is more appropriate to compare the expected data with data set aside and not used to reconstruct the structure. In this case the term is called "*R*free". For a correctly reconstructed structure, one expects the ratio *R*/*R*free to be >80%.

Of course atoms in a crystal also have thermal motion. We can estimate the extent of their motion by looking at their electron density and, indeed, crystallography assigns a value that describes the extent to which the electron density is spread out to each atom. This value, called the "temperature factor" or "Debye–Waller factor" or *B* factor is given by:

$$B = 8\pi^2 U^2$$

where *U* is the mean displacement of the atom (in Å), so high B factors indicate greater uncertainty about the actual atom position. For example, for $B = 20$ Å^2:

$$U = \frac{1}{\pi}\sqrt{\frac{B}{8}}\text{Å} = \frac{1}{\pi}\sqrt{\frac{20}{8}}\text{Å} \cong \frac{1}{3.14}\sqrt{2.5}\text{Å} \cong 0.5\text{Å}$$

and the uncertainty about the position of the atoms is 0.5 Å. Values of 60 or greater may imply disorder (for example, free movement of a side-chain or alternative side-chain conformations). As expected, atoms with higher *B* factors are often located on the surface of a protein whereas the positions of the atoms in the internal well packed core of the protein are less uncertain (Figure 1.16). Finally, the occupancy value for an atom represents the fraction of expected electron density that was actually observed in the experiment.

Nuclear magnetic resonance (NMR) is another very useful technique for determining the structure of macromolecules. This technique is based on the observation that several nuclei (e.g. H, ^{13}C, ^{15}N) have an intrinsic magnetic moment. If we place a concentrated homogeneous solution of a protein (or nucleic acid) inside a

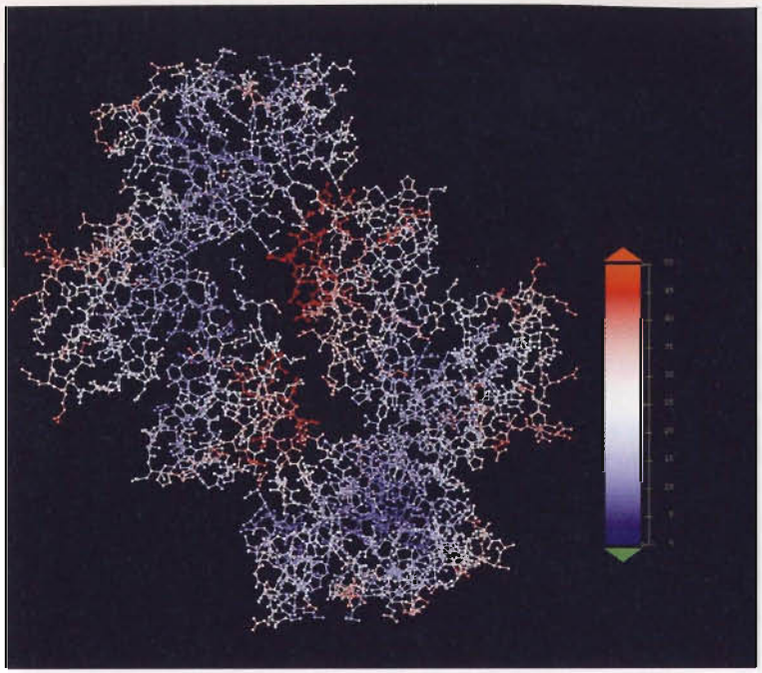


Figure 1.16 A protein structure colored according to the B-factor of its atoms. The color scheme is such that atoms with high B-factors are red and those with low B-factors blue.

very powerful magnetic field, the spin of the nuclei will become oriented in the direction of the external field. By applying radio-frequency magnetic fields to the sample, we can measure the energy absorbed at the frequency corresponding to the jump between two allowed spin orientations. Each atom has a characteristic resonance which depends on its structure, but it is also affected by the surrounding atoms. These subtle absorbance differences between the same atom in different environments make it possible to identify which resonance corresponds to which of the protein atoms.

If two atoms are close in space, magnetic interactions between their spins can be measured. The intensity of the interaction decays rapidly with the distance between the atoms (it is a function of r^{-6}, where r is the distance). This effect can be exploited to map short distances between interacting atoms. The result of the experiment is a set of lower and upper limits for the distance between pairs of atoms (constraints). If the number of constraints is sufficient, there will be a finite number of possible conformations of the protein compatible with the data. The more constraints we are able to measure, the more similar to each other will these structures be (Figure 1.17).

The number of constraints in an NMR experiment is strongly dependent on the flexibility of the protein and of its regions in solution: if a given region is very mobile, it will be very difficult to identify the neighbors of its atoms because they

will not spend enough time next to each other. In such cases, we cannot measure the interactions but we recover very valuable information about the intrinsic mobility of the protein structure.

Question: How do I evaluate the quality of an NMR structure and how does it compare with X-ray structure?

»NMR structures are usually reported with *rmsd* values (the square root of the average sum of the squared distances between corresponding atoms, see later) for the various structures compatible with the data. The lower the *rmsd*, the

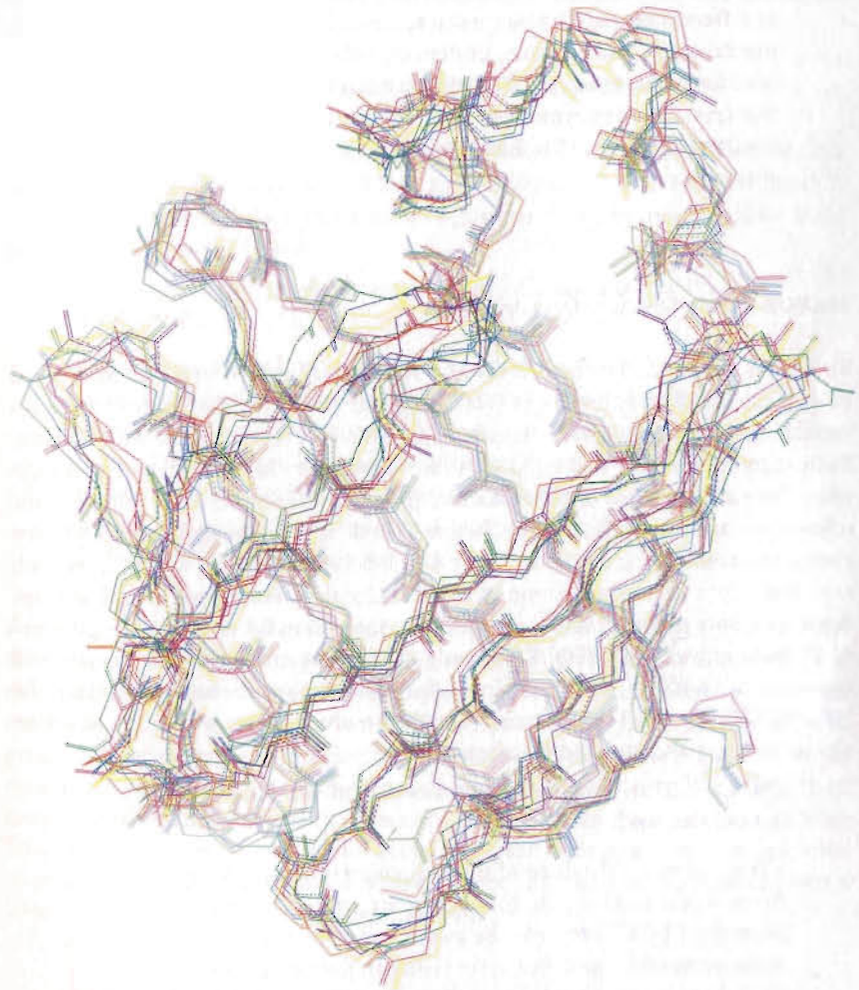


Figure 1.17 Several NMR-derived structures of the chicken fatty acid-binding protein. Note that the exposed regions are less well defined than the central core of the protein.

higher the accuracy of the measurements. The answer to the second part of the question depends on whether the question is posed to a crystallographer or to an NMR spectroscopist! There is no clear and definite answer because the two experiments give different, albeit related, information.«

Question: Does the crystal structure of a protein reflect its "true" native and functional structure?

»This question is often asked. Several lines of evidence point to a positive answer – structures of the same protein solved by both X-ray crystallography and NMR, or solved independently in different crystal forms, are the same within the experimental error. Furthermore, protein crystals are full of solvent (and for this reason very fragile) and it has often been shown that crystallized enzymes can function inside the crystal; they are therefore deemed to have the correct native functional structure.«

1.5 The PDB Protein Structure Data Archive

Structures determined by both X-ray and NMR are deposited in a data base called PDB. X-ray structure entries consist of a single structure; for NMR entries there is a variable number of structures, usually approximately 20, compatible with the data. Each entry is uniquely identified by a four-letter code. In the first part of a PDB entry there are the name of the molecule, the biological source, some bibliographic references, and the R and Rfree factors. There is also information about how chemically realistic the model is, i.e. how well bond lengths and angles agree with expected values (the values found in small molecules). For a good model, average deviations from expected values should be no more than 0.2 Å in bond lengths and 4° in bond angles. The SEQRES records contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule studied, whereas the HELIX, SHEET and TURN records list the residues where secondary structure elements begin and end and their total length.

Question: Where can I find the sequences of all the proteins of known structure?

»There is also a database of the sequences of known structures, usually called pdb, containing the sequences extracted from the SEQRES records. Be aware, however, that even if some parts of the protein are not visible in the electron density map, because they are too mobile or because the protein was partially degraded in the experiment, their sequence will still be included in the SEQRES field. In other words the sequence

corresponds to that of the studied molecule, not necessarily to the part of the molecule the structure of which is contained in the entry. The database of sequences of known structure called ASTRAL only includes the sequence of the part of the molecule that has been experimentally determined.«

After this initial part of the file, the actual coordinates are listed in records identified by the keyword ATOM. These include a serial number for the atom, the atom name, the alternative location indicator, used when the electron density for the atom was observed in two positions, the chain identifier, a residue sequence number and code, the x, y, and z orthogonal coordinates for the atom, the occupancy, and the temperature factor. For example the record:

```
ATOM 1281 N GLY Z 188A 29.353 66.969 17.508 1.00 28.84
```

describes the nitrogen atom of a glycine unit with residue number 188 and residue code A. The coordinates are $x = 29.353$, $y = 66.969$, $z = 17.508$. The occupancy is 1.00 (i.e. complete) and the B factor is 28.84 (corresponding to an uncertainty in the position of this atom of 0.6 Å.

Question: Which is the minimum occupancy of atoms reported in a PDB file?

»There is no lower limit to the value of the occupancy for an atom. It can be 0 if the position of an atom was guessed on the basis of the positions of the surrounding atoms. Be aware that none of the widely used structure-visualization packages highlights them automatically. It is always advisable, if one is working on a particular region of a protein, to verify the B factor and occupancy of its atoms.«

It is worth briefly describing the residue number and code, because these are often the cause of much frustration when trying to use a PDB file: the residue number is not necessarily consecutive. For example, trypsin is synthesized as a longer molecule the first 15 amino acids of which must be enzymatically removed to produce the active protein. The first residue number in the 3PTI entry for trypsin is indeed 16. A common numbering scheme is occasionally used for a family of evolutionarily related proteins, and in such circumstances the residue numbering follows the scheme. If one of the proteins of the family contains amino acids inserted among the commonly accepted numbering, the residue code is given a letter. In the 3TPI entry, for example, we find:

```
ALA 183
GLY 184A
TYR 184
LEU 185
GLU 186
```

GLY 187
GLY 188A
LYS 188

For NMR structures the headers do not, of course, include the R factors and the resolution. The ATOM fields are quite similar, the B factor is usually set to 0 and the sections referring to each of the models are included between the records MODEL and ENDMDL.

1.6 Classification of Protein Structures

Protein structures can be classified according to their similarity, in terms of secondary structure content, fold, and architecture. There are a few widely used

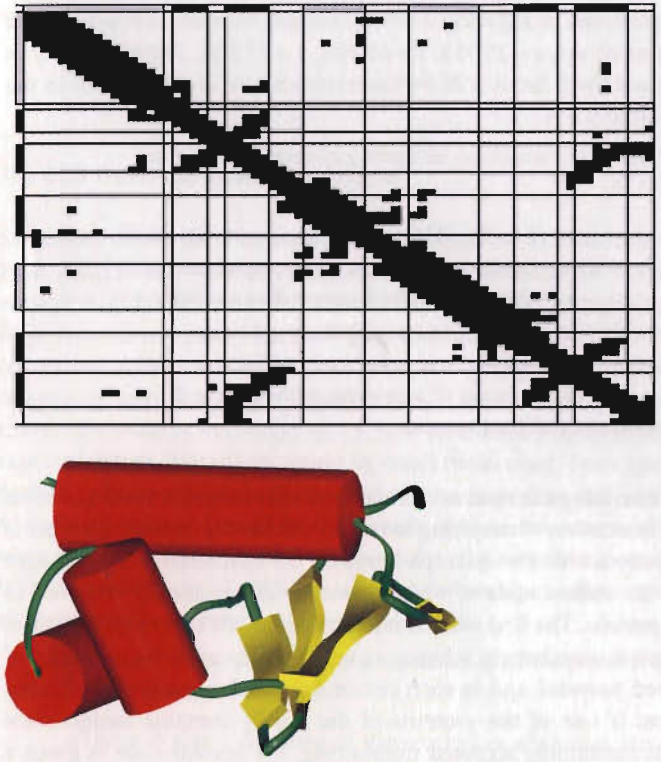


Figure 1.18 A distance map for the domain shown at the bottom (PDB code: 1RUN). The secondary structure of the protein is shown in the first line and first column of the matrix. Black and gray regions correspond to beta strands and alpha helices, respectively. Filled cells correspond to distances shorter than 6Å.

classifications of protein structures which are extremely useful for navigating through the protein structural space. These are collected and made available to the community via Web servers and differ in the method used to obtain the classifications.

FSSP is a classification method based on comparison of the "distance matrices" of proteins. These are an alternative representation of protein structures (Figure 1.18) obtained by filling a matrix. Each row and each column of a matrix represent an amino acid and each cell contains the distance between the amino acid in the row and the amino acid in the column in the protein structure. Given two proteins, we can compare their distance matrices and derive a structural superposition between their atoms, i.e. the superposition that minimizes the distance between corresponding pairs of atoms. The resulting structural distance between the two proteins, defined as the root mean square of the average sum of the squared distances, is used by FSSP to cluster the known structures and to classify them.

CATH is another classification of protein structures based on use of a different algorithm to compute structural similarity. In this classification the two distance matrices that are compared contain the vectorial distance between pairs of atoms, rather than the scalar one. CATH provides a hierarchical classification of the structures, identifying four levels of similarity – Class, Architecture, Topology, and Homology. The Class is defined on the basis of the predominant type of secondary structure (all alpha, all beta, alpha and beta, and domains with little or no secondary structure). The Architecture describes the overall shape of the domain structure as determined by the orientations of the secondary structures ignoring the connectivity between the secondary structures. It is assigned on the basis of visual inspection of the proteins and of literature data. The Topology level depends on the structural distance between proteins, and evolutionarily related proteins are grouped at the Homology level, on the basis, as we will see, of sequence-based methods.

Finally, SCOP is another classification with a hierarchical organization including Class, Fold, Superfamily, and Family levels. The main Class types in SCOP are all alpha, all beta, alpha plus beta, and alpha/beta. A protein is assigned to one of the classes according its predominant secondary structure. The other classes include multi-domain, membrane, and cell surface proteins and peptides, small proteins, peptides, designed proteins, and low-resolution structures. The second level of classification, Fold, includes proteins with similar topological arrangements for which an evolutionary relationship cannot be identified, the third level (Superfamily) includes proteins that are believed to share a common ancestor. Proteins related by an unambiguous evolutionary relationship are grouped at the Family level. The classification in SCOP is essentially manual, although some automatic pre-processing is used to cluster clearly similar proteins.

None of these classifications is intrinsically better than any other and they usually agree with each other.

1.7 The Protein-folding Problem

The stability of each possible conformation of an amino acid chain depends on the free energy change between its folded and unfolded states:

$$\Delta G = \Delta H - T\Delta S$$

where ΔG, ΔH, and ΔS are the differences between the free energy, enthalpy, and entropy, respectively, of the folded and unfolded conformations. The enthalpy difference is the energy associated with atomic interactions within the protein structure (dispersion forces, electrostatic interactions, van der Waals potentials, and hydrogen bonding that we will describe in more detail later) whereas the entropy term describes hydrophobic interactions. Water tends to form ordered cages around non-polar molecules, for example the hydrophobic side-chains of an unfolded protein. On folding of the polypeptide chain, these groups become buried within the protein structure and shielded from the solvent. The water molecules are more free to move and this leads to an increase in entropy that favors folding of the polypeptide.

Question: What does an unfolded protein looks like?

»Although we generally assume that the unfolded chain is in a random coil conformation, i.e. that the angles of rotation about the bonds are independent of each other and all conformations have comparable free energies, the reader should be aware that, in reality, unfolded proteins tend to be less disordered and more compact than ideal random coils, because some regions of the polypeptide can interact more favorably with each other than with the solvent.«

In a cell proteins are synthesized on ribosomes, large molecular assemblies comprising proteins and ribonucleic acid molecules. Special adaptor molecules, tRNA molecules, recognize a triplet of bases on the messenger RNA, which in turn has been synthesized by following instructions contained in the genome, and adds the appropriate amino acid to the nascent chain. The synthesis of an average protein takes approximately a minute; the time required for folding, i.e. for achieving the "working" native structure, is comparable. Some slow steps of the reaction, for example formation of disulfide bonds, are accelerated by specific enzymes. Other proteins are also involved in the folding process and their role is either to protect the nascent protein chain (shielding the hydrophobic regions that are exposed to solvent before folding occurs) or to provide a more protected environment for folding; there is no evidence that anything but the amino acid sequence determines the native protein structure in vivo.

In the nineteen-sixties the American chemist Christian Anfinsen and his coworkers performed a series of seminal experiments demonstrating that the native

conformation of a protein is adopted spontaneously or, in other words, that the information contained in the protein sequence is sufficient to specify its structure. The enzyme selected by Anfinsen for the experiment was ribonuclease A (RNase A), an extracellular enzyme of 124 residues with four disulfide bonds (Figure 1.19). As already mentioned, these are covalent bonds arising as a result of oxidation of the sulfhydryl (SH) groups of the side-chains of two cysteines, when they are close to each other. The result is an S–S bond between their sulfur atoms. In Anfinsen' s experiment, the S–S bonds were first reduced to eight –SH groups (by use of mercaptoethanol, a reducing agent with the chemical formula HS–CH_2–CH_2–OH); the protein was then denatured by adding urea in high concentration (8 Molar). (The urea molecule enhances the solubility of nonpolar compounds in water and therefore reduces the strength of the stabilizing hydrophobic interactions that hold the protein structure together.) Under these conditions the enzyme is inactive and becomes a flexible random polymer. In the second phase of the experiment the urea was slowly removed (by dialysis); the –SH groups were then oxidized back to S–S bonds. We expect that if the protein is able to assume its

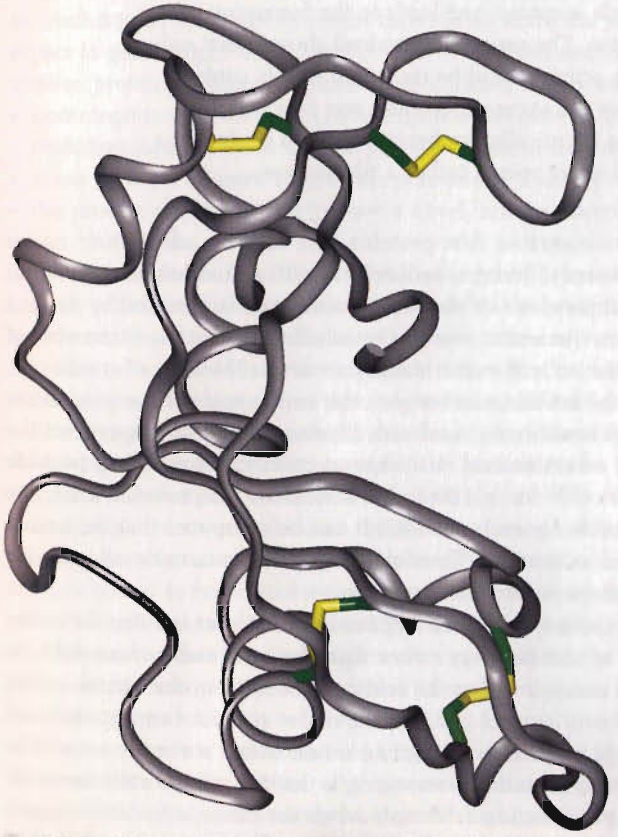


Figure 1.19 The structure of ribonuclease A (PDB code: 1AFK). Note the four disulfide bridges.

correct tertiary structure, the correct pairs of cysteines are close to each other so that the correct disulfide bonds form and the protein regains its activity. Indeed, the refolded protein regained more than 90% of the activity of the untreated enzyme.

Question: How can we be sure that the protein was really unfolded after adding urea and mercaptoethanol?

»Anfinsen and co-workers also performed a control experiment to demonstrate that RNase A was completely unfolded in 8 Molar urea. RNase A was reduced and denatured as above, but in the second phase the enzyme was first oxidized to form S-S bonds, and only afterwards was the urea removed. If the protein is really in a random conformation in 8 Molar urea, it is likely that the cysteines are in different relative positions in different molecules and will randomly pair giving rise to scrambled sets of disulfide bonds. Because there are eight cysteine residues in ribonuclease, there are $7 \times 5 \times 3 \times 1 = 105$ different ways of forming disulfide bonds, only one of which is correct and leads to the formation of a functional enzyme. The experiment indeed showed that only about 1% of the activity could be recovered in this control experiment. Later the same experiment was successfully repeated using a chemically synthesized protein chain, i.e. a protein that had never seen a cell or a ribosome.«

These experiments demonstrated that proteins can, indeed, adopt their native conformation spontaneously, but immediately raised a fundamental problem known as the Levinthal paradox – if the same native state is achieved by various folding processes both in vivo and in vitro, we must conclude that the native state of a protein is thermodynamically the most stable state under "biological" conditions, i.e. the state in which the interactions between the amino acids of the protein are the most energetically favorable compared with all other possible arrangements the chain can assume. But an amino acid chain has an enormous number of possible conformations (at least 2^{100} for a 100-amino-acid chain, because at least two conformations are possible for each residue). It can be computed that the amino acid chain would need at least ~2^{100} ps, or ~10^{10} years to sample all possible conformations and find the most stable structure.

Levinthal concluded that a specific folding pathway must exist and that the native fold is simply the end of this pathway rather than the most stable chain fold. In other words, Levinthal concluded that the folding process is under kinetic rather than under thermodynamic control and that the native structure corresponds not to the global free energy minimum but rather to one which is readily accessible. The hypothesis underlying Levinthal' s reasoning is that the energetically favorable contacts that stabilize the structure arise only when the chain is folded or nearly folded. In other words, the protein chain must first lose all its entropy (being locked in a given conformation) and, only when the correct conformation is reached, can

the entropy loss be compensated by the gain in enthalpy. A wealth of literature addresses the Levinthal paradox and we will not dwell on the details here, except to say that, in general, the paradox can be solved by thinking of the folding process as a sequential process in which the entropy decrease is immediately or nearly immediately compensated by an energy gain and that, in this hypothesis, the time-scale computed for the folding process approximates that observed in nature.

1.8 Inference of Function from Structure

The Structural Genomics Initiatives promise to deliver between 10 000 and 20 000 new protein structures within the next few years and, as we will see in this book, many more protein structures will be modeled. The challenge is obviously to exploit this large amount of structural data to predict the functions of these proteins. Proteins sharing a common evolutionary origin (homologous proteins) have similar structures, as we will see shortly, and, occasionally, proteins that do not seem to share an evolutionary relationship might turn out to share the same topology. One can expect to gain insight into a protein' s function from analysis of other, structurally similar, proteins. There are at least three difficulties to be overcome in this process:

- homologous proteins might have originated by gene duplication and subsequent evolution and therefore have acquired a different function;
- some folds are adopted by proteins performing a variety of function; and, finally,
- the protein of interest might have a novel, not yet observed, fold.

What can we learn from the analysis of a protein structure? We can certainly identify which residues are buried in the core of the protein and which are exposed to solvent. The structure will also tell us the quaternary structure of the protein – the structure observed in the crystal is often that which is biologically active, although there are exceptions that might create difficulties.

The presence of local structural motifs with functional roles can be detected by analyzing the structure. For example, the presence of a helix–turn–helix motif suggests that the protein binds DNA. Two alpha helices intertwined for approximately eight turns with leucine residues occurring every seven residues are the dimerization domains of many DNA-binding proteins. A motif in which a zinc atom is bound to two cysteines and two histidines separated by twelve residues is called a zinc finger and is found in DNA and RNA binding proteins. Other shorter and non-contiguous local arrangements can be identified and associated with a function, for example the arrangement of serine, histidine and aspartate in serine proteases (Figure 1.3).

When no known local functional motif can be detected, it is still possible to analyze clefts on the surface of the protein (in more than 70% of proteins the largest cleft contains the catalytic site) and highlight the presence of amino acid side-chains that are likely to be involved in catalytic activity. Biochemical knowledge can help us to postulate a catalytic mechanism.

For non-enzymes, the problem is much harder to solve. Detecting the protein–protein interaction sites is very difficult and there is not yet a completely satisfactory method, although analysis of the hydrophobicity of the surface in conjunction with automatic learning approaches is leading to some success.

When other members of the evolutionary family are known, analysis of the conservation and variability of amino acids facilitate estimation of the functional importance of different parts of the structure. Any approach used to detect function from structure has a major limitation, however: the molecular function of a protein does not tell us very much about its biological role. If we predict a protease activity for an enzyme, even if we can identify the likely substrate, we are still left with the question of its biological role, because these enzymes participate in many processes, from digestion to blood coagulation, from host defense to programmed cell death.

It should also be mentioned that, recently, more and more proteins (called moonlight proteins) have been found to perform more than one function, often totally unrelated to each other. This might be frustrating, but should not be surprising – there is no reason evolution should not take advantage of different surface regions of proteins to endow them with different activities.

Another significant proportion of proteins seem to be intrinsically disordered and assume their native structure only when they meet and bind with their partners (natively unfolded proteins).

> *Question: Is the property of being disordered functionally important ?*
>
> »The property is often evolutionarily conserved and is, therefore, deemed to be functionally relevant. The reason might be that their flexibility enables these proteins to bind several targets or to provide a large interacting surface in big complexes. This might also be a clever way of engineering high specificity but low affinity. A large interaction surface usually confers both properties but, if the protein has to expend energy for folding before binding, specificity can still be achieved without large affinity. Other explanations can be invoked for this behavior, for example the lifetime of an unfolded protein in a cell is probably shorter and this can provide a regulatory mechanism. More simply, there is no reason why evolution should select against these proteins, because selective pressure acts on the function of the protein and is not concerned with what the protein does when not involved in its functional interactions, assuming it does not have a deleterious effect on the cell.«

The existence of moonlight and natively unfolded proteins makes the problem of inferring the function of novel proteins even more complex and, indeed, this is one of the fields that is attracting more attention at the present. It is easy to predict that many new more powerful methods will be developed in the near future, taking advantage both of the wealth of data that is being accumulated and of novel approaches. The problem is somewhat recent – before the start of structural genomics initiatives, the determination of the structure of a protein was usually the final step of its characterization and was aimed at understanding the details of its functional mechanism or interactions rather than to infer its biological function from scratch. Only recently are we facing the challenge of having an available structure and no functional information.

1.9
The Evolution of Protein Function

In 1859 Darwin published "The Origin of Species", a book that laid the foundation of evolutionary theory. The careful observations he made during his travels led him to realize that the taxonomy of species could be explained by postulating gradual changes occurring generation after generation and to propose that changes might result in competitive advantage for the organism as members of a population better fitted to survive leave more offspring. The traits of successful individuals then become more common, whereas traits that do not increase, or even reduce, the fitness become rarer or disappear altogether. Evolution, therefore, acts to transform species in the direction of better fitness for the environment. Darwin also had the intuition that even traits that do not, by themselves, confer any selective advantage, might become predominant in a population if they attract the preference of sexual partners. At about the same time Mendel discovered that the traits of the partners are not blended in the offspring – on the contrary, specific characters are sorted and inherited. The foundations for a molecular theory of evolution only needed identification of the material carrying the characteristics, i.e. the DNA, and this happened approximately half a century later.

At the time it did seem surprising that a simple molecule such as DNA, which is, after all, only a polymer comprising a limited set of different nitrogen-containing bases (only four, as it happens) each attached to a sugar and a phosphate group, could explain the diversity between an amoeba and a man. Only fifty years later, however, the diffraction data collected by Rosalind Franklin enabled James Watson and Francis Crick to build a structural model of DNA. The structure of this molecule immediately suggested how DNA can be replicated and copied. This is probably the only example in history in which knowledge of the structure of a macromolecule has immediately provided information about a novel functional mechanism.

What remained to be understood was how the DNA could code for proteins, i.e. what was the code linking the four-character alphabet of a DNA molecule with the twenty-letter alphabet of proteins. The path that led to unraveling this code was much harder than the previous steps, it took years of study and experiments to obtain the genetic code (Table 1.2), i.e. the correspondence between each triplet of bases of the DNA and the coded amino acid. With rare exceptions, the genetic code

is universal, it is used by bacteria, plants, animals, more proof, if needed, of the theory of evolution.

Table 1.2 The genetic code.

	U	C	A	G	
	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
	Leu	Ser	**STOP**	**STOP**	A
	Leu	Ser	**STOP**	Trp	G
	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

A story related to the genetic code is very instructive for highlighting that biology is a complex field and that beautiful elegant theories might fall short of describing how evolution has shaped life. Francis Crick was one of the contributors to the long and labor-consuming path that led to the discovery of the genetic code. In 1957, however he devised the following solution to the problem of mapping the four DNA bases to the twenty amino acids. We have four bases and twenty amino acids. There are only 16 possible combinations of two characters out of an alphabet of four (4^2), therefore a coding system that associates a pair of bases with each amino acid could only encode sixteen amino acids. If we use a triplet of bases to code for an amino acid, then we have 64 (4^3) possible combinations. The problem is how to map 64 triplets to twenty amino acids. Let us assume that only a subset of the 64 possible triplets codes for amino acids and that they are such that, when two are placed next to each other, only one "reading frame" can be meaningful. For example, if the codons CGU and AAG code for an amino acid, then none of the triplets GUA and UAA can be coding sequences, so if the DNA contains the sequence CGUAAG there is no ambiguity in how it should be read and translated.

In this hypothesis, none of the triplets AAA, CCC, GGG, and UUU can be coding, because if they were the sequences AAAAAA, CCCCCC, GGGGGG and UUUUUU would be ambiguous, so we are left with 60 codons. Next, only one of the codons that are cyclic permutations of each other can be coding. Let us consider

the codons ACG, CGA and GCA. Ambiguities arise if more than one of these is used. For example, if we use ACG and CGA, the sequence ACGACG is ambiguous. This implies that we can only select one codon every three; we are therefore left with only 20 out of the 60 codons. Crick and colleagues did realize that, apart for its elegance and simplicity, there was no other support for this hypothesis and indeed they wrote: "We put it forward because it gives the magic number – 20 – in a neat manner and from reasonable physical postulates." The theory was very elegant and equally wrong, demonstrating that evolution selects a working alternative, and does not seem to be interested in elegant minimal design!

Each cell of an organism, with rare exceptions, carries a complete copy of the genetic material. Bacteria usually have only one copy of the genetic material organized as a circle made of double-stranded DNA. More complex species have cells with nuclei and reproduce sexually. Most of the DNA in such organisms resides in the cell nucleus and is arranged in several chromosomes that occur in pairs. One member of each pair comes from the mother and one comes from the father. Some DNA is also stored in separate organelles within the cell, called mitochondria and chloroplasts.

Darwin viewed evolution in terms of the genealogical relationships among species or major groups of organisms over a long time span. The impressive progress in molecular biology enables us to study evolution in molecular terms, by looking at the change in the genetic make-up of a population and at the differences between species in terms of difference in their DNA sequence.

Replication, either in the process of creating new somatic cells (mitosis) or in the process of creating germ cells (meiosis), is extremely accurate, and there are several mechanisms to ensure its fidelity; errors are inevitable, however. Environmental factors, for example high-energy radiation, can, moreover, cause random damage to the DNA molecule. Mechanisms exist for repairing the damage, but sometimes they introduce errors. These can be of two types – replacements of DNA bases by others or deletions or insertions of any number of bases. A base replacement may or may not affect a protein sequence. The change may occur in an intron or in another region of the DNA that does not code for a protein. When it occurs in a protein–coding region, the replacement might lead to a codon that is translated into the same amino acid as the original, because of the redundancy of the genetic code. Alternatively, an amino acid residue in the original protein may be replaced by a different amino acid in the mutated protein (missense mutation) or the mutation can involve a stop codon. If a codon for an amino acid residue is changed to a stop codon, the protein will be terminated prematurely and will usually be non-functional (nonsense mutation) whereas if a stop codon mutates into a codon for an amino acid residue the translation continues, elongating the amino acid chain until the next stop codon is encountered.

Large insertions or deletions in the coding regions of a protein almost always prevent production of a useful protein. Short deletions or insertions in a coding region of any number of bases other than a multiple of three usually have a drastic effect – they cause a shift in the reading frame during translation, resulting in a meaningless change in the amino acid sequence in the C-terminal direction from

the point of mutation. When the insertion or deletion involves multiples of three bases, it does not affect the sequence of the protein outside the site of the insertion or deletion and may or may not affect its function.

A gene, or a whole chromosomal region, might be duplicated, leading to a situation in which two copies of the same gene are present. If there is no selective pressure, the two copies may evolve independently – one copy may continue to code for the protein performing the original function whereas the other may evolve by mutation into an entirely different protein with a new function. New combinations of existing genes are occasionally produced at the beginning of meiosis when the chromatids, or arms, of homologous chromosomes break and reattach to different chromosomes (crossing-over). It is easy to see how these mechanisms can account for variability within a species and differences between different species.

One can also speculate about the mechanisms by which new species arise. A species is defined as a set of individuals that, in the wild, would mate and produce fertile offspring. A new species can therefore originate when some individuals, for whatever reason, do not mate with the rest of the population for a sufficient length of time. These individuals can follow a different evolutionary path that might result in genetic incompatibility with the original group. This can be because of physical separation between groups of individuals, or acquisition of different lifestyle, or spreading of the individuals over a huge geographical range. Study of evolution and of the relationships between species and their proteins is of paramount importance in modern biology; most of what we can infer about the function and structure of biological elements comes from analysis of their differences and similarities with the corresponding elements in different species. The possibility of comparing the sequence of entire genomes has also resulted in the possibility of highlighting which parts of the genome are under evolutionary pressure and are, therefore, deemed to be functionally important. It is not surprising that a plethora of tools and theories has been developed to highlight evolutionary relationships, some of which will be described in this book in the appropriate context. Here we will review some elements of the terminology that are commonly used.

Phylogeny is an inferred pattern of evolutionary relationships between different groups of organisms. Usually we depict a phylogeny as a rooted tree in which the length of the branches is proportional to the divergence time and each leaf represents a species. We also use a tree representation to indicate the evolutionary relationships between genes or proteins. In this representation each leaf is a gene or a protein and the lengths of the branches are proportional to the accumulated changes between the molecules. It should be kept in mind that, although molecular trees derived from protein sequences are related to phylogenetic trees, the former refer to the observed difference, for example in functional regions, and do not, therefore, necessarily relate to a proper phylogenetic tree, because different evolutionary pressure can result in different rates of evolution for different genes or proteins. For example, the rate of mutation of hemoglobin is approximately one change per site every billion years whereas fibrinopeptides can accumulate nine time this number of mutations in the same period of time.

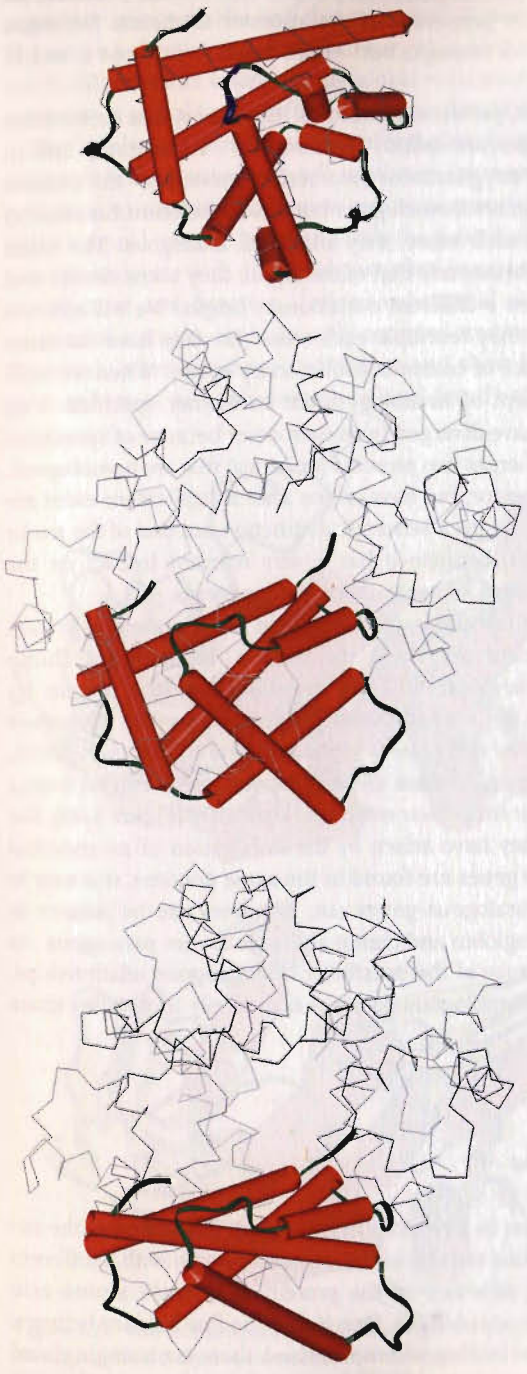


Figure 1.20 Three homologous chains: myoglobin, and alpha and beta globin. Note that these three structures are similar and evolutionarily related, but they are paralogous, i.e. they have arisen by gene duplication.

Sometimes molecular trees are unrooted, i.e. they depict differences, but make no hypothesis about the location or properties of the ancestor elements. For some applications, such a tree is good enough, but – once again – it is not a tool to derive evolutionary times.

Two elements (whether genes, proteins, anatomical structures) that derive from a common evolutionary ancestor are called "homologous". In anatomy and in protein analysis homology does not guarantee common functionality – the anterior wings of a bat and a human arm are homologous but have a different function. If two anatomical parts resemble each other, they are called analogous. The usual example here is the eye of vertebrates and that of the squid, they seem similar and have a similar function, but have a different evolutionary origin. We will call two protein structures analogous if they resemble each other (i.e. they have the same topology) but there is no evidence of common evolutionary origin. When we refer to genes or proteins, the concept of homology must be further specified. Two proteins (or genes) believed to have diverged from each other because of speciation events are called orthologous whereas two proteins (or genes) that are homologous, i.e. derived from a common ancestor, but have arisen after a duplication event are called paralogous. This is by no means a semantic distinction but one of the major issues in protein bioinformatics, because it has a very relevant impact on the prediction of the biological function of newly discovered proteins.

Let us use an example to illustrate the issue. Myoglobin is a monomeric protein the function of which is to store oxygen in the muscle. Human and chimp myoglobin are orthologous – they descend from the same ancestral protein via speciation and they also have the same function. Hemoglobin, the tetrameric oxygen transporter is composed by two pairs of alpha and beta chains. Myoglobin, alpha hemoglobin, and beta hemoglobin are all homologous, they descend from a common ancestor, as is apparent from their structural similarity (Figure 1.20), but they are paralogous, because they have arisen by the duplication of an ancestral globin gene. If two homologous genes are found in the same genome, it is easy to see that they are paralogous. Paralogous genes can, however, also be present in different genomes – human myoglobin and chimp alpha globin are paralogous. As illustrated by this example, because of the possibility of paralogous relationships, the finding that two genes are homologous does not necessarily imply they share the same function.

1.10
The Evolution of Protein Structure

If a base-substitution event occurs in a protein-coding region of a genome, the net effect can be the substitution of one residue of the encoded protein with a different one. What is the effect on the structure of the protein of a single amino acid replacement? There are only two possibilities. One is that the fine balance between the gain and loss of free energy of folding is compromised, there is no single global energy minimum for the new sequence, and it does not fold any more. Because

proper folding is required for function, the most likely outcome is that the organism is not viable and the mutation is not propagated in the population. The second alternative is that the energy landscape of the new sequence changes, but it still contains a free energy global minimum and the corresponding native structure is still able to perform the same function as the original protein.

How likely is it that the new conformation is very different from the original? Statistics and physics both tell us this is extremely unlikely and that the most probable outcome is that the new sequence assumes a structure very similar to that of the original protein. In other words the substituted amino acid is accommodated into the structure with only local perturbation and without dramatic global changes in structure and function. Indeed substantial changes in protein architecture, because of a single, evolutionarily accepted mutation, have not yet been observed. Therefore, when residue substitutions and short insertions/deletions accumulate in members of an evolutionarily related family of proteins, they will cause local

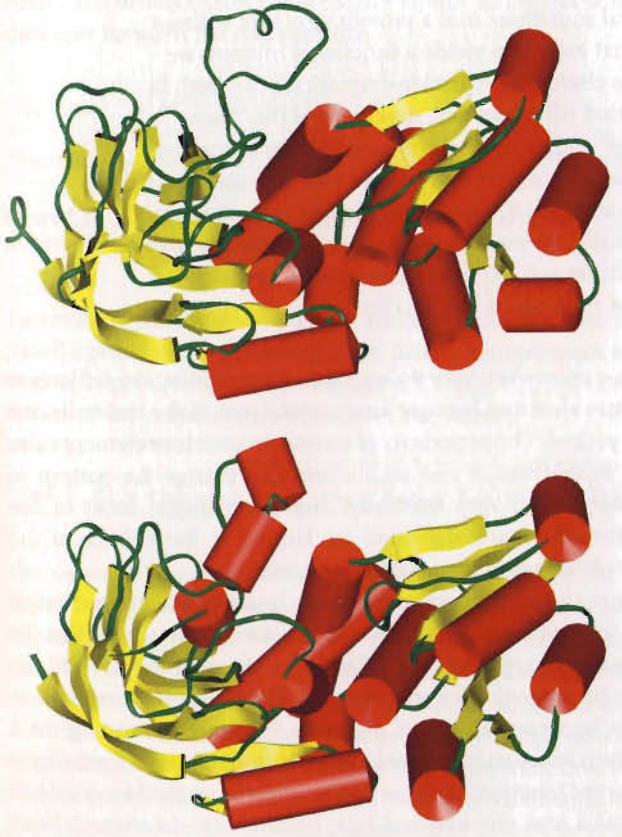


Figure 1.21 Two homologous proteins sharing 30% sequence identity: alcohol dehydrogenase from horse liver (PDB code: 1YE3) and *Acinetobacter calcoaceticus* (1F8F).

structural perturbations without affecting the general shape, or topology, of the protein. Of course, the greater the number of mutations (or, equivalently, the further the proteins are in the evolutionary scale), the larger will be the difference between the protein structures. We can quantify this qualitative observation by measuring the relationship between the different sequences of homologous proteins and their structural divergence, as we will see in the next section. We will merely mention here that it is accepted that pairs of evolutionarily related proteins sharing at least 30% sequence identity have a similar fold (Figure 1.21).

> *Question: Is the sequence-to-structure relationship limited to naturally evolved proteins or does it reflect an intrinsic property of amino acid sequences?*
>
> »It is important to understand that the sequence–structure relationship already described occurs when there is the requirement, as in natural evolution, that each step of the evolutionary path must be functional. If we were to artificially introduce several mutations into a protein structure without guaranteeing that each step yields a functional mutant, we might be able to change its structure dramatically. Indeed, in 1994 two scientists (Creamer and Rose) issued the "Paracelsus Challenge" – they offered a prize of $1000 to the first group to successfully convert one protein fold into another while retaining at least 50% sequence identity with the original fold (this challenge was named after Paracelsus, the 16th century Swiss alchemist). At least three groups have published reports in response to this challenge.«

Insertions and deletions, even when they do not cause frame shifts, are difficult to accommodate in a protein structure because a substantial part of the molecule, the internal core, is tightly packed. The periodicity of secondary structure elements also implies that insertion or deletion of one amino acid can change the pattern of interaction of the whole region very markedly. Not surprisingly, most of the observed amino acid insertions and deletions are located at the surface of the protein structure and outside secondary structure elements.

Fusion of two or more initially independent genes leads to the production of multidomain proteins with new combinations of functions in a single protein. In eukaryotes, this process is thought to be facilitated by the presence of introns (intervening sequences in genes that are not coding). These represent regions in which genes can easily be recombined – if one exon from a gene coding for a protein region with a given function is inserted into an intron region of a gene for a protein carrying a different function, the new hybrid protein might be capable of both functions and serve a new physiological role. It should be mentioned, however, that there is no evidence that exons preferentially encode structural or functional units.

1.11
Relationship Between Evolution of Sequence and Evolution of Structure

To analyze the relationship between sequence and structural divergence in quantitative terms, we must define a measure of distances in sequence and structure space. Several unsolved problems connected with this issue will be discussed later in this book. For the moment let us assume we know how to find the correspondence between the amino acids of two evolutionarily related proteins that reflects their evolutionary history. In other words, let us assume that, given two proteins, we can construct a matrix such as that shown in Figure 1.22 in which the first row contains the amino acid sequence of the first protein, possibly with inserted spaces, and the second contains the amino acid sequence of the second protein, again possibly including spaces. Two amino acids in the same column are assumed to originate from the same amino acid of the ancestor protein. The spaces represent insertion and deletion events. Given this correspondence, called alignment, we can define the distance in sequence space simply as the fraction of amino acids that is different between the two proteins.

Sequence 1	...	M	Q	D	G	T	S	R	F	T	C	R	G	K	–	–	P	I	H	H	F
Sequence 2	...	L	S	E	G	N	H	K	L	S	C	R	H	D	Q	G	P	V	N	D	Y

Figure 1.22 Alignment of two fragments of the sequences of the proteins shown in Figure 1.21.

To measure distance in structure space, we use the root mean square deviation (*rmsd*) between corresponding atom pairs after optimum superposition. In practice, we apply the rigid-body translation $T = (T_x, T_y, T_z)$ and rotation $R = (R_x, R_y, R_z)$ to one of the proteins that minimizes the value:

$$rmsd(T, R) = \min_{T,R} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - R_x x_i' + T_x)^2 + (y_i - R_y y_i' + T_y)^2 + (z_i - R_z z_i' + T_z)^2\right]}$$

The set of corresponding atom pairs should, once again, be that which reflects the evolutionary correspondence between amino acids. Obviously, when we are superimposing two proteins with different sequences, we can only use atoms that are common between any two protein structures, for example the Cα, or the atoms of the backbone, or the backbone plus the Cβ for all amino acids except glycine.

As already discussed, peripheral parts of the proteins can undergo local rearrangements that can be quite substantial. If there are insertions and deletions it is obviously impossible to compute the *rmsd* values for the inserted residues, but even if this is not so, the fact that the superposition procedure minimizes the sum of the squares of the deviations means the *rmsd* is dominated by the most divergent regions and, if we include them, the changes in the more conserved regions would be masked. This leads to the need to superimpose separately the conserved "cores"

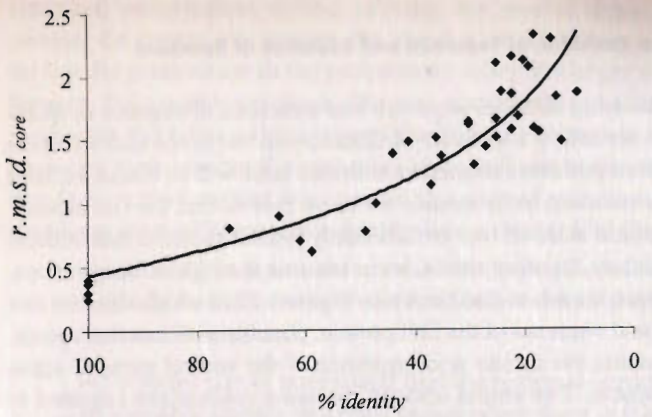


Figure 1.23 Relationship between sequence identity and structural similarity. The plot is obtained using the same set of proteins originally analyzed by Lesk and Chothia.

of evolutionarily related proteins and, therefore, calls for a definition of the core. Many empirical procedures are commonly used. Chothia and Lesk, for example, propose the following: given two related proteins, one first superimposes the main chain atoms of corresponding elements of the secondary structure and then continues to add residues at either ends of the elements until the distance between the alpha carbons of the last added residue deviates by more than 3Å. Next, one jointly superimposes these "well fitting" regions and calculates the resulting *rmsd*.

Now that we know how to measure sequence distance and structural divergence, we can investigate the relationship between them. In a seminal paper Chothia and

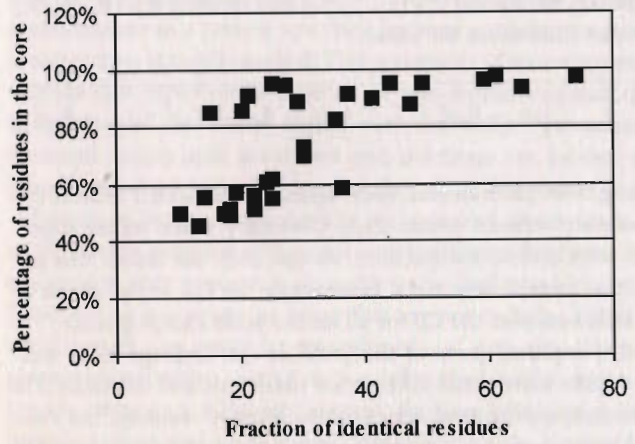


Figure 1.24 Relationship between sequence identity and the extent of the common structural core between pairs of homologous proteins. (Data from the original Chothia and Lesk analysis on thirty-two pairs of proteins).

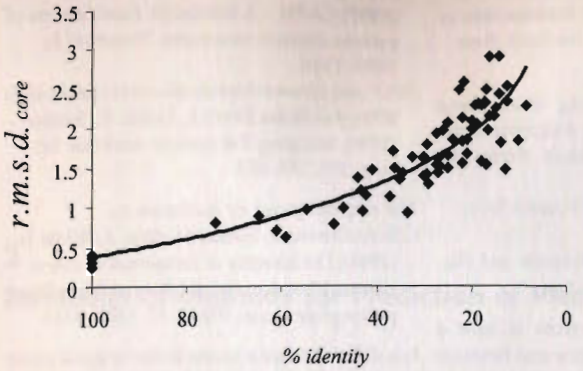


Figure 1.25 Relationships between sequence identity and structural similarity. The plot was obtained by using a larger set of proteins than in Figure 1.23, but the trend is essentially the same.

Lesk selected thirty-two pairs of homologous proteins of known structure, identified the common core within each pair with the procedure described above, and computed the *rmsd* values between the core of each pair as a function of the sequence identity between the two protein sequences (Figure 1.23). Their conclusions were that, as sequences diverge, the extent of the common core between two homologous proteins decreases. The common core contains almost all the residues when pairs of closely related proteins (with sequence identity >50%) are considered; when residue identity drops below 20% the structures might diverge quite substantially and the core can contain as little as 40% of the structure (although in some cases it can include most of the structure) (Figure 1.24). They found that the relationship between the structural divergence of the core and the sequence identity was in accordance with the equation:

$$rmsd_{\text{core}} = 0.40e^{\frac{(100 - \%\text{identity})}{100}}$$

Although the original analysis by Chothia and Lesk was limited to 32 pairs of proteins only, a relationship with very similar parameters was obtained when the analysis was repeated using a much larger sample (Figure 1.25).

Suggested Reading

These two books guide the reader through a fascinating tour of protein architecture and show how the shape of a protein is linked to its function:

A.M. Lesk (2001) Introduction to Protein Architecture, Oxford University Press

A.M. Lesk: (2004) Introduction to Protein Science: Architecture, Function, and Genomics, Oxford University Press

Another excellent book that describes the principles of protein structure, with examples of key proteins in their biological context, is:

C.-I. Branden, J. Tooze (1999) Introduction to Protein Structure, 2nd edn, Garland, New York

Readers interested in learning more about crystallography and nuclear magnetic resonance spectroscopy can consult these two seminal books:

J. Drenth (1994) Principles of Protein X-ray Crystallography, Springer
K. Wüthrich (1986) NMR of Proteins and Nucleic Acids, John Wiley and Sons

Every biochemistry book contains at least a chapter dedicated to the structure and function of proteins. Readers can read the relevant chapters from:

J.M. Berg, L. Stryer, J. L. Tymoczko (2002) Biochemistry, 5th edn, W. H. Freeman
D.L. Nelson, M. M. Cox (2004) Lehninger Principles of Biochemistry, 4th edn, W. H. Freeman
D. Voet, J. Voet (2004) Biochemistry, 3rd edn, Wiley

The original paper describing the PDB data archive (http://www.pdb.org) is:

F.C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi (1977) The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. J. Mol. Biol. 112, 535–542

Structural classification of proteins, SCOP (http://scop.mrc-lmb.can.ac.uk/scop), is described in:

A.G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540

CATH (http://www.biochem.ucl.ac.uk/bsm/cath/) C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton

(1997) CATH – A hierarchic classification of protein domain structures. Structure 5, 1093–1108

FSSP (http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html) L. Holm, C. Sander

(1996) Mapping the protein universe. Science 273, 595–602

The original paper by Anfinsen is:

C.B. Anfinsen, E. Haber, M. Sela, F. White Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. PNAS 47, 1309–1314

It is difficult to gain access to the original paper in which Levinthal described the paradox that took his name:

C. Levinthal (1969). In: P. Debrunner, J. C. M. Tsibris, E. Munck (Ed.) Mossbauer spectroscopy in biological systems, University of Illinois, Urbana, IL, pp. 22–24

However, several more recent papers describe the reasoning explained in the paper.

The worldwide initiatives in structure genomics are described at http://www.rcsb.org/pdb/strucgen.html where a good collection of informative background information about this project is also available.

I would also recommend reading "The Origin of Species" by Charles Darwin, because of its outstanding historical and scientific interest.

The paper by Francis Crick and colleagues, describing their theory on the comma free genetic code is:

F.H.C. Crick, J. S. Griffith, L. E. Orgel (1957) Codes without commas. PNAS USA 43, 416–421

Finally, the analysis of the relationship between sequence and structural similarity in proteins by C. Chothia and A. M. Lesk can be found in:

C. Chothia, A. M. Lesk (1986) The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823–826

2
Reliability of Methods for Prediction of Protein Structure

2.1
Introduction

Imagine a scenario in which a biologist has obtained the amino acid sequence of a previously unknown protein and is trying to gather as much information as possible on the molecule. The first thing he or she will do is to access the internet and try to find and use the many tools available, including those, which we will discuss later, that produce three-dimensional models of the protein of interest. Undoubtedly, if more than one method is used, the results will not coincide. Occasionally, the discrepancies between the models obtained will be small, but for more difficult structures even the same method will present the user with more than one solution to the problem. The issue is, therefore, how do we select the best model, i.e. the one we will use as a theoretical framework to investigate further our biological system?

Given more than one model, we will certainly try to see which of the proposed structures is more consistent with available experimental information. Before investing our time in this enterprise, however, we would like to know which of the results are expected to be more reliable, by looking at the quality of the results that each method has produced in the past. Furthermore, as we will discuss later, it is possible that none of the available methods can produce a model sufficiently good for our purposes.

This chapter is devoted to description of how the accuracy of a method is evaluated and what should we look at when using it.

Two methods are usually used to estimate of the quality of a method. In the first we select a set of cases for which the answers, for example the three-dimensional structures, are known, pretend we do not know them and verify how similar the results of the method are to the real experimental answer. In the second, we use the method to predict the structure of a protein that is not yet known, but that will soon be elucidated, and, when the data are available, compare the predicted and observed features. Both strategies have advantages and disadvantages, which we will discuss; before this, however, we must address the problem of how we can measure the difference between the predicted and experimental structures.

Prediction of a protein structure can be limited to prediction of the location of secondary structure elements, to production of an alignment to a protein of known structure, or to a fully fledged three-dimensional model of the protein.

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

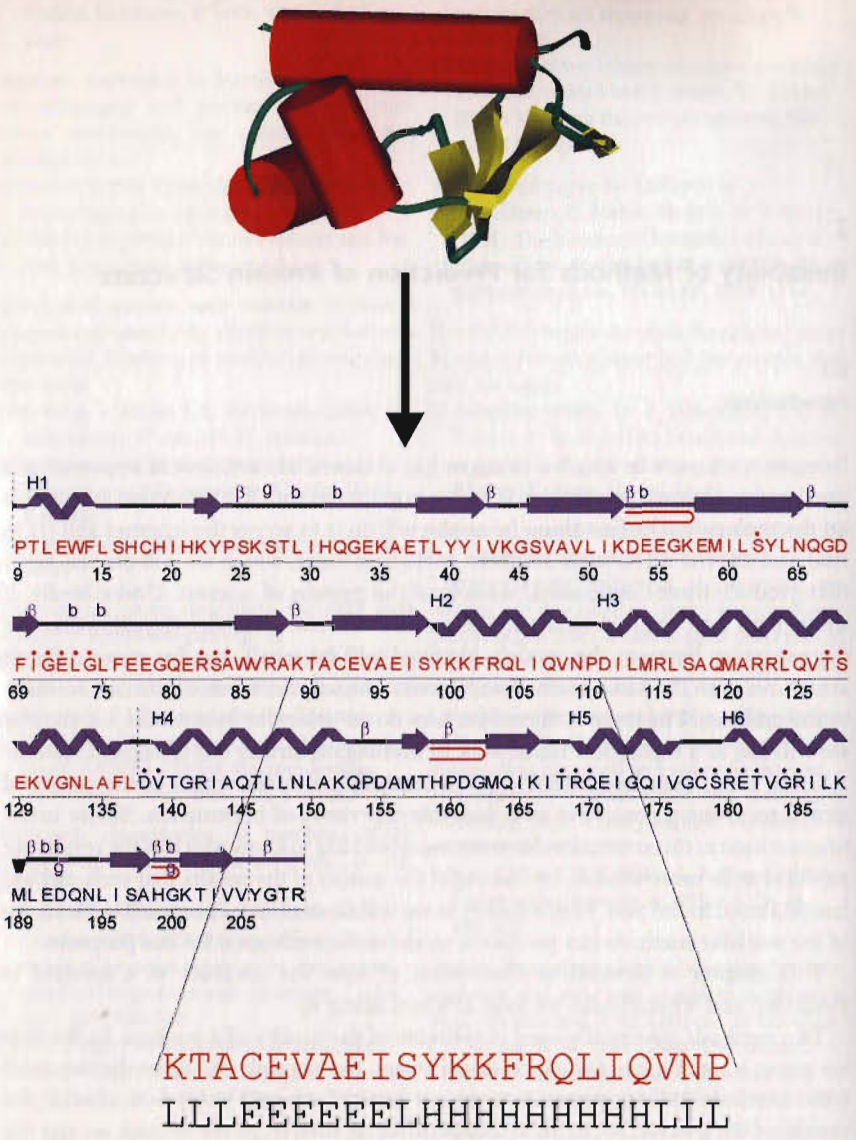


Figure 2.1 The secondary structure, predicted or experimental, of a protein can be encoded as a string. Usually only alpha helices (H) and beta strands (E) are considered. 3(10) helices are included in the helical definition, and every other conformation is indicated with L (loop). The figure shows the three-dimensional structure of an SH3 domain, a very useful summary of its properties that can be found at the site http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/, a database providing an overview of every macromolecular structure deposited in the Protein Data Bank, and the string encoding the secondary structure of a fragment of the protein.

2.2
Prediction of Secondary Structure

We can envisage the results of a secondary structure prediction method as a string of characters representing, for example, residues predicted to be in alpha helices, beta strands, and irregular regions. This string must be compared with the experimentally determined secondary structure of the protein which can also be encoded as a string (Figure 2.1).

It is, however, a fact that elements of regular structure in proteins are not as regular as we would like them to be, especially at their termini. For example, an alpha helix is defined both by its ϕ and ψ backbone angles and by its typical hydrogen-bond pattern. It is quite common that the last residue of the helix does not form all its hydrogen bonds or even that, if the helix has a bend, some internal residues lose one or two hydrogen bonds (Figure 2.2). Similarly, in beta sheets

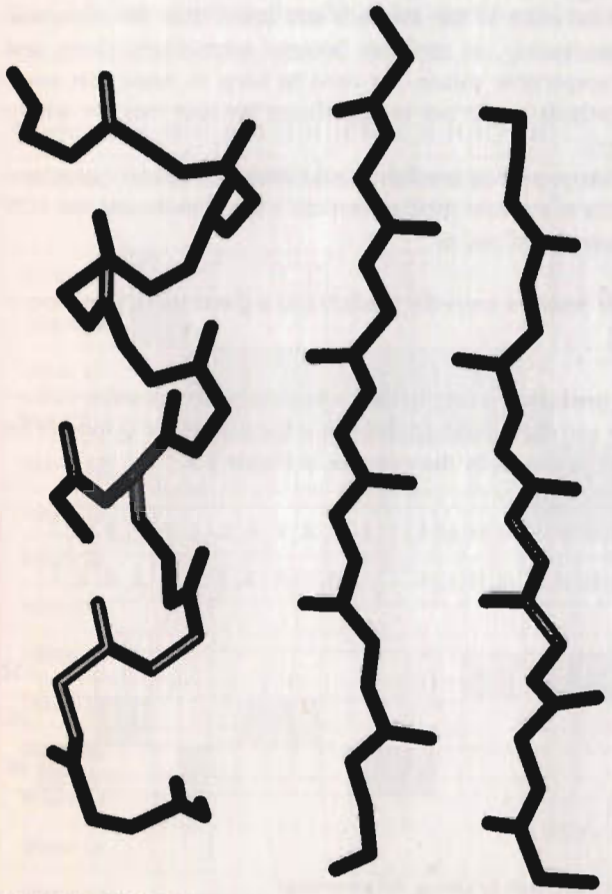


Figure 2.2 Helices and strands contain irregularities highlighted by the absence of the standard hydrogen-bond pattern.

there might be bulges (i.e. a region of irregularity in a beta sheet, in which the normal pattern of hydrogen bonding is disrupted by insertion of one or more extra residues) or the strands might move slightly further away from each other at their ends and lose the regular pattern of hydrogen bonds that characterizes them (Figure 2.2).

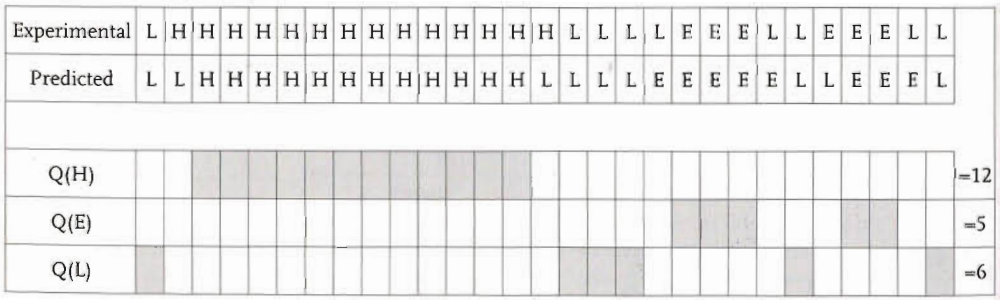
Several methods are widely used to assign the limits of the secondary elements; the two most commonly applied are probably DSSP and STRIDE. The first is essentially based on the pattern of main chain hydrogen-bonds between amino acids; the second uses empirical energetic criteria to define the presence and/or absence of hydrogen-bonds and also takes into account the values of the ϕ and ψ angles. Other methods are based on Cα–Cα distances or on the curvature of the peptide backbone. Each of these methods has its drawbacks. That, in general, these methods do not all agree on secondary structure assignment for a given protein is simply because there is no right answer.

Whereas a few years ago this was a negligible problem in secondary structure prediction, because the accuracy of the methods was lower than the aforementioned experimental uncertainty, as methods become increasingly clever and accuracy reaches very respectable values, we need to keep in mind that small differences between methods might not be significant because they are within experimental uncertainty.

Given the two strings representing predicted and observed secondary structure, we usually use two terms to evaluate their agreement – the *Q* index and the *SOV* score. The Q index is simply defined as:

$Q = 100 \times$ (Number of residues correctly predicted in a given state)/(Total number of residues)

In secondary structure prediction most methods assign only three possible states – helix, strand, and other and the *Q* index (called *Q*3) is the sum of the *Q* indexes for each of the three states, as shown by the example in Figure 2.3.



Q3 = 12/29 + 5/29 + 6/29 = 0.79

Figure 2.3 The Q3 value is computed by adding the percentage of residues correctly predicted to be in the helical, strand, or loop conformation.

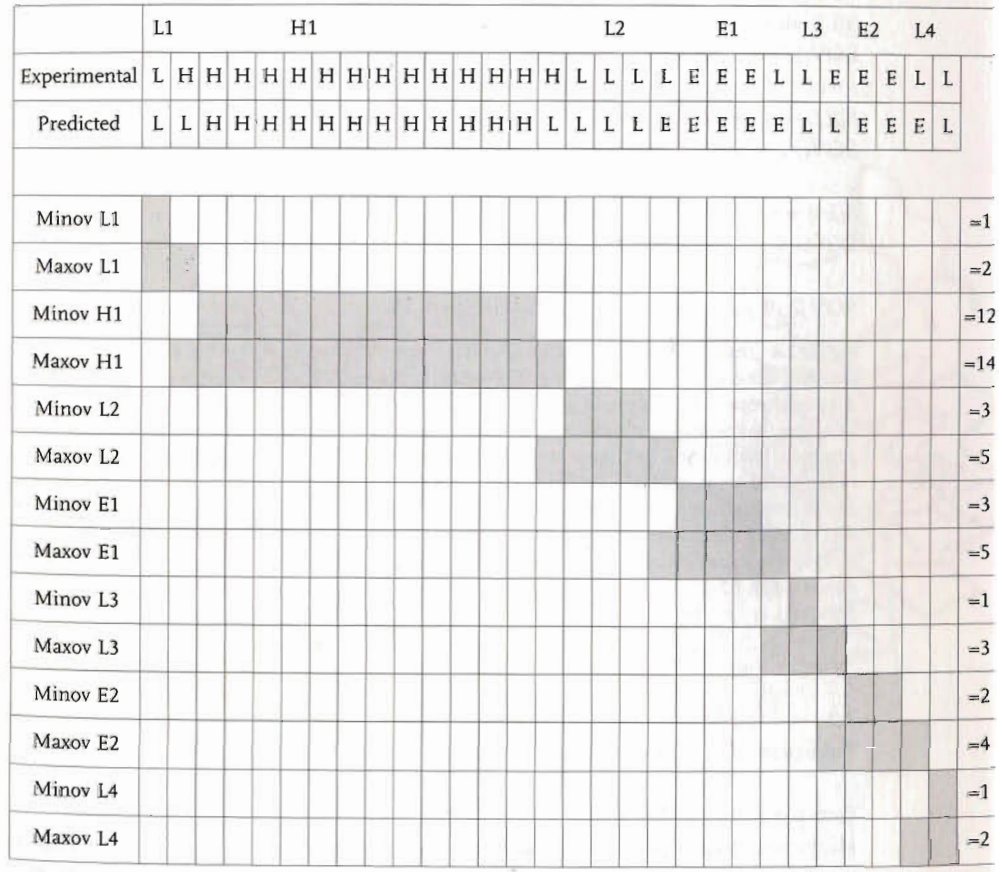
The *SOV* measures the overlap between segments, rather than the accuracy of the prediction of the state of a single residue. For one of the states, for example the helices, we define:

$$SOV(i) = \sum_i \frac{Minov(S1_i, S2_i) + \Delta(S1_i, S2_i)}{Maxov(S1_i, S2_i)} \times len(S1_i)$$

$Minov\ (S1_i, S2_i)$ is the length of the regions in which two corresponding secondary structure segments overlap (two elements are defined as corresponding if they have at least one residue in common). *Maxov* is the length of the region in which either has residues in the secondary structure and:

$$\delta(S1_i, S2_i) = min\left((Maxov(S1_i, S2_i) - Minov(S1_i, S2_i), Minov(S1_i, S2_i), int\left(len\frac{S1_i}{2}\right), int\left(len\frac{S2_i}{2}\right)\right)$$

The *SOV* for the three states is computed as the sum of the three *SOV* measures divided by the sum of the lengths of the experimental segments (Figure 2.4).



Experimental:
len(L1) = 1; len(H1) = 14, len(L2) = 4; len(E1) = 3; len(L3) = 2; len(E2) = 3; len(L4) = 2
Predicted:
len(L1) = 2; len(H1) = 12, len(L2) = 4; len(E1) = 5; len(L3) = 2; len(E2) = 3; len(L4) = 1

δ(L1) = int(min(2−1,1,int(2/2),int(1/2))) = int(min(1,1,1,0.5)) = 0
SOV(L1) = ((1+0)/2)*1 = 0.5

δ(H1) = int(min(14−12,12,int(12/2),int(14/2))) = int(min(2,12,6,7)) = 2
SOV(H1) = ((12+2)/14)*14 = 14

δ(L2) = int(min(5−3,3,int(4/2),int(4/2))) = int(min(2,3,2,2)) = 2
SOV(L2) = ((3+2)/5)*4 = 4

δ(E1) = int(min(5−3,3,int(5/2),int(3/2))) = int(min(2,3,2.5,1.5)) = 1
SOV(E1) = ((3+1)/5)*3 = 2.4

δ(L3) = int(min(3−1,1,int(2/2),int(2/2))) = int(min(2,1,1,1)) = 1
SOV(L3) = ((1+1)/3)*2 = 1.34

δ(E2) = int(min(4−2,2,int(3/2),int(3/2))) = int(min(2,2,1.5,1.5)) = 1
SOV(E2) = ((2+1)/4)*3 = 2.25

δ(L4) = int(min(2−1,1,int(2/2),int(1/2))) = int(min(1,1,1,0.5)) = 0
SOV(L4) = ((1+0)/2)*2 = 1

SOV = (0.5+14+4+2.4+1.34+2.25+1)/29 = 0.88

Figure 2.4 The Figure shows how to calculate the *SOV* value for a prediction. One *SOV* value is computed for each segment of secondary structure. In the example there is one helical segment, two strands, and three loops. For the helix, *maxov* is 14, *minov* is 12 and the lengths of the predicted and observed helical segment are 14 and 12, respectively. Thus delta, the integer minimum between ((*maxov-minov*), *minov*, 14/2, 12/2) is 2. The resulting value for *SOV*(H1) is 14. For the strands, we must compute an *SOV* value for each segment. The first has *maxov* = 5, *minov* = 3, and the lengths of the experimental and predicted regions are 3 and 5, respectively. This implies that the value of delta for this segment is 1 and its *SOV* is 2.4. For the second strand, *maxov* = 4, *minov* = 2, delta = int(min(2,2,1,1) = 1 and therefore *SOV* = 2.25. At the end, the *SOV* values for each segment are summed and divided by the length of the sequence.

2.3 Prediction of Tertiary Structure

Comparison of the structures of two different proteins requires solution of both the alignment and superposition problems – we must find which pairs of atoms we want to superimpose and then find the translation and rotation of one of the two

molecules that minimizes the *rmsd* between the sets of paired atoms. When comparing a model and its experimental structure the alignment problem is trivial, because the two structures have the same sequence, but this does not mean there are no other problems. Consider the example shown in Figure 2.5. On the left is the superposition between a structure and its respective model obtained using all the Cα atoms; on the right the superposition is computed using about a half of the Cα atoms.

The inspection of the figure might convince the reader that the superposition shown on the right is more representative of the overall quality of the model, because it shows how well the packed core of the structure is predicted, and does not take into account the deviations in peripheral parts of the structure. The superposition shown on the right can also be very informative for the end user if, as in this example, it involves the set of atoms that are part of the active site of the molecule. The fraction of superimposed structure and the *rmsd* values are, obviously, correlated and this implies that, given two different predictions for the same structure, we need to compare the *rmsd* values for an equivalent fraction of superimposed atoms and, at least in principle, the *rmsd* values for the biologically important regions should be considered more relevant. A useful way of obtaining a qualitative impression of the relative accuracy of two models is to plot the *rmsd* values as a function of the fraction of superimposed atoms. In the example shown

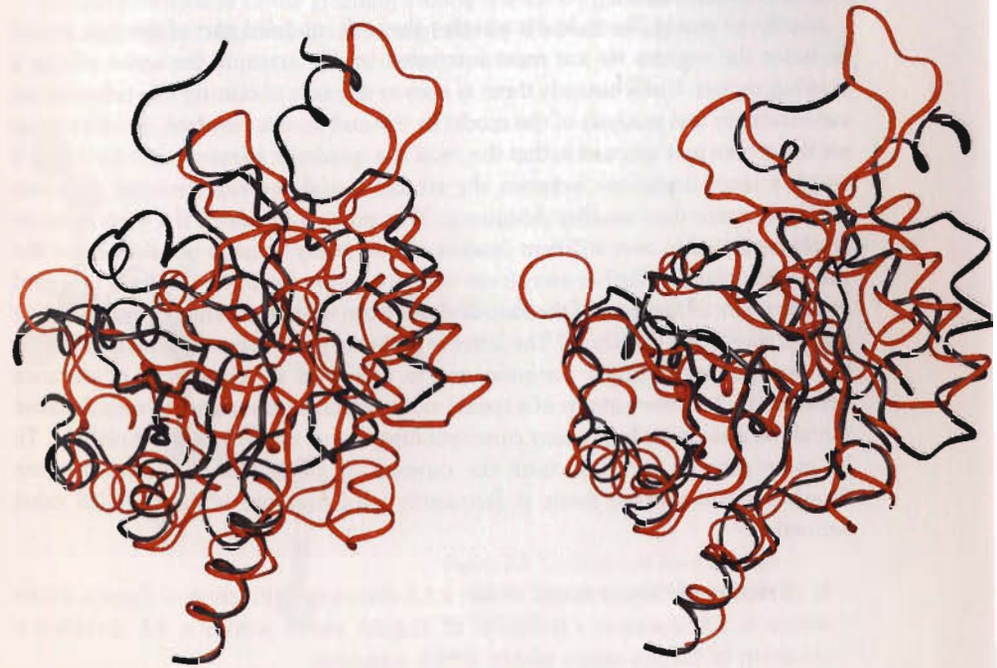


Figure 2.5 The result from structural superposition depends both on the number of superimposed atoms and the *rmsd* value.

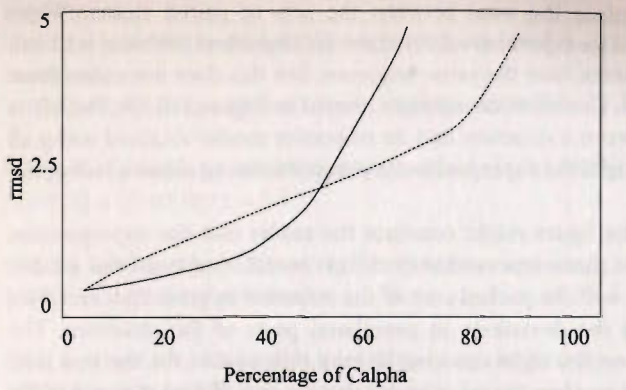


Figure 2.6 Given a model and an experimental structure, the *rmsd* values for the best superposition can be computed by superimposing 5%, 10%, 15%, etc. of the model and target structures. Two models can be compared by plotting the *rmsd* as a function of the fraction of superimposed atoms. In the example, the model represented by the continuous line fits better than the model represented by the dotted line, when approximately one half of the structure is considered. The situation is inverted when larger fragments are superimposed.

in Figure 2.6, it is clear that the model indicated by the continuous line is closer to the experimental structure for about one half of the structure whereas the other has a better overall quality.

Ideally we would like to know whether the well modeled part of the first model includes the regions we are most interested in, for example the active site or a binding region. Unfortunately there is no way of easily obtaining this information automatically and analysis of the model by the end user is required. Another issue we must take into account is that the *rmsd* is a quadratic measure and therefore it weights large deviations between the structure and the experimental structure relatively more than smaller deviations. This poses a problem: if a loop is incorrectly predicted in two different models, do we really want to penalize more the model that placed it farther away from the right answer or do we just want to regard both answers as incorrect if the loop deviates from the correct answer by more than a given acceptable threshold? The latter is probably more reasonable and, indeed, it has become increasingly common not to use *rmsd* as a measure of distance between the backbone atoms of a model and a structure, but rather to set a distance threshold and count how many corresponding atoms are within the threshold. To be more general, we may count the number of atoms within several distance thresholds and average them. A frequently used measure is the GDT-TS value defined as:

¼ ((Fraction of Calpha atoms within a 1 Å distance) + (Fraction of Calpha atoms within a 2 Å distance) + (Fraction of Calpha atoms within a 4 Å distance) + (Fraction of Calpha atoms within an 8 Å distance))

Of course, different distance thresholds can be used.

The GDT-TS measure gives an estimate of the correctness of the overall structure but does not consider, for example, how well the side chains are modeled. The measure that is more often used for assessing the quality of side-chain predictions is based on comparison of the dihedral angles of the side chain, usually limiting the calculation to, at most, the first two angles, called χ_1 and χ_2 (Figure 2.7) and computing the fraction for which the deviation is lower than a given threshold, usually 30°.

The next issue we must consider is the quality and possible uncertainties in the experimental structure. In different cases, one might wish to exclude some subsets of atoms from the superposition, for example:

- Atoms with high B factor. If the model is reasonably close to the experimental structure, it is advisable not to consider atoms the B factors of which are very high, because this indicates their position is not well defined.
- Solvent-exposed side chains. Exposed side chains, especially if they are long, are usually mobile.
- Regions involved in crystal contacts. Regions involved in crystal contact between different molecules in the crystal might have a conformation different from that in solution.
- Side-chains with ambiguous chi values. Occasionally rotation of 180° around some side-chain bonds might be undetectable by X-ray crystallography, for example the rotation around the Cα alpha and Cβ beta bond of valine (affecting the computed value of the χ_1 angle) around the Cβ Cγ gamma bond of histidine and asparagine (affecting χ_2), or around the Cγ–Cδ bond of asparagine (affecting χ_3).

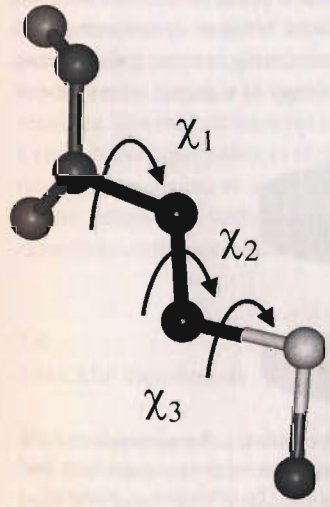


Figure 2.7 Definition of the side-chain χ angles. The amino acid shown here is methionine.

2.4 Benchmarking a Prediction Method

Benchmarking is the process of comparing methods, services, and products in a specifically defined and strictly controlled environment. In structure prediction, this translates into selecting a set of cases with known answers and using them to evaluate the accuracy of a prediction method. Most structure prediction methods are based on heuristics, they use data that must be estimated using a set of proteins for which the structure is known (training set). The accuracy of the predictions on the training set is expected to be higher than on a different test set. Because homologous proteins have similar structure, the quality of the predictions will also be higher on proteins homologous with proteins in the training set. Therefore the test set should contain proteins that share no significant sequence similarity with the training set and, if the test is to be balanced, both training and test sets should have a similar distribution of structure classes and types.

Other biases might be more subtle and difficult to avoid. The length, composition, and cellular localization of a protein might have an effect on its structure. Many methods use information derived from the evolutionary history of the protein under examination, so the number of known homologous proteins in the evolutionary family might be another piece of information that should be taken into consideration when selecting the appropriate test and training sets.

Because the number of proteins of known structure is limited, rather than using separate training and test sets, structure prediction methods often employ cross-validation or jack-knife techniques. In a jack-knife test of N proteins, one protein is removed from the set, the parameters are estimated from the remaining $N-1$ proteins, and then the structure of the removed protein is predicted and the accuracy of the prediction is measured. This process is repeated N times by removing each protein in turn. If the estimation step is very time-consuming, a more limited cross-validation can be performed by splitting the sample into M subsets, where $M < N$.

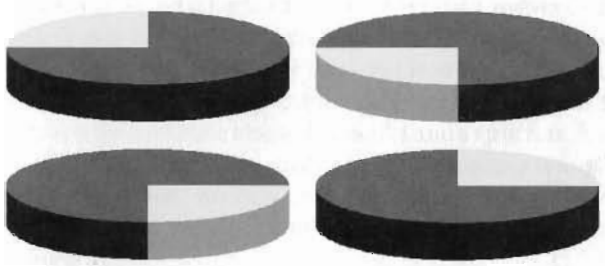


Figure 2.8 Benchmarking consists in selecting a subset of the data to derive parameters and use the remaining data to test the method. The data are split into a test set (dark gray) and a training set (light gray). The training set is used to derive the results and these are tested on the test set. Next, a different non overlapping set of test data is selected, and their properties are predicted using parameters derived from the remaining data. The procedure is repeated until all data have been used once as test sets. The accuracy of the method is the average accuracy on the different test sets.

Parameters are divided from $(M-1)N$ proteins and tested on the remaining N proteins. This process is repeated M times, once for each subset (Figure 2.8).

2.5 Blind Automatic Assessments

Benchmarking is not always possible, especially if the training procedure is very time consuming. It is, furthermore, useful to evaluate different methods on the same test set to be able to compare their results. For methods that can be run automatically, the solution is to set up an automatic system that collects the results of different methods as soon as a new protein structure is determined, and therefore before any method had a chance to use it in the training set. This has the added advantage that the methods can use all available data in their training set and are not restricted to deriving parameters from a more limited training set.

EVA is one server that performs this useful service to the community. Every day, EVA downloads the newest protein structures from the PDB, extracts the sequences for every protein chain, and sends them to each prediction server registered for the experiment. The results collected are then evaluated and made public. EVA covers several methods that predict solvent accessibility, secondary structure, and complete three-dimensional modeling. The proteins used in the experiment are such that no pair of them has more than 33% identical residues over more than 100 residues aligned.

Livebench is another continuous benchmarking server, but it limits itself to the evaluation of three-dimensional models of proteins not sharing a significant sequence similarity (and therefore deemed to be non-homologous) to any protein of known structure. Every week, new entries in the PDB database are submitted to participating servers. Every week the results are collected and evaluated. In this experiment a target is skipped if it is shorter than 100 residues or longer than 500 residues. The evaluation uses only the Cα positions of the models. It first performs a rigid-body superposition of the model and the structure and then computes the maximum number of atoms within 3 Å after superposition.

The results of both servers are publicly available via the Internet and are extremely useful tools that should be consulted before using any prediction server.

2.6 The CASP Experiments

EVA and Livebench can be used to evaluate the performances of automatic servers, but these are not the complete scenario of prediction methods. First, there is at least the hope that human intervention, with exploitation of data from the target protein, can improve the models. Second, there are methods that cannot easily be automated, either because they are in early stages of their development, or because they require manual evaluation of the results.

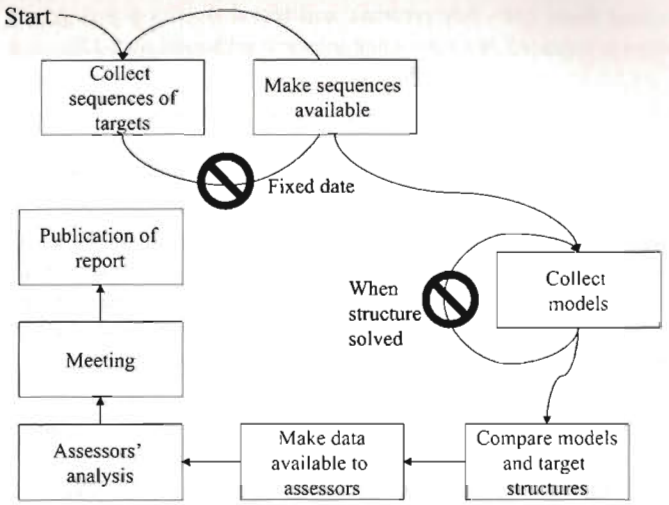


Figure 2.9 The CASP experiment runs every two years. In the spring, approximately, targets are collected from experimenters working on the resolution of their structure. The sequences are made available to predictors who can submit predictions until the structure is solved. Numerical comparison of models and targets is performed by a group of scientists led by John Moult and Krzystof Fidelis. The data are then passed to thee assessors, chosen by the community on the basis of their expertise, who analyze the data and try to derive general conclusions about the state of the art in the prediction field. In approximately December of the same year, predictors, assessors, and organizers convene in a meeting to discuss the results and, later, publish the final reports in the scientific journal *Proteins: Structure, Function and Bioinformatics.*

In 1994 John Moult proposed a world-wide experiment named CASP (critical assessment of techniques for protein structure prediction) aimed at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused. The organization of the experiment is depicted in Figure 2.9.

Crystallographers and NMR spectroscopists who are about to solve a protein structure are asked to make the sequence of the protein available, with a tentative date for release of the final coordinates. Predictors produce and deposit models for these proteins (the CASP targets) before the structures are made available. CASP also tests publicly available servers on the same set of targets providing a unique opportunity to verify the effectiveness of human intervention in the modeling procedure. A panel of three assessors compares the models with the structures as soon as they are available and tries to evaluate the quality of the models and to draw some conclusions about the state of the art of the different methods. The task is divided among assessors in such a way that one looks at models of proteins that share significant sequence similarity with a protein of known structure, one looks at those sharing a significant structural similarity, but no clearly detectable sequence similarity with proteins of known structure, and the third evaluates all the

remaining models. It is expected that the first set of targets is predicted using a technique called comparative modeling and the second using fold-recognition methods; because the experiment is run blind, however, i.e. the assessors do not know who the predictors are until the very end of the experiment, it is entirely possible that different techniques are used by different groups for the same target. We will use this division in the rest of this book to describe the various approaches to structure prediction, although, as we will discuss, there is substantial overlap between them.

The results of the comparison between the models and the target structures are discussed in a meeting where assessors and predictors convene, the conclusions are made available to the whole scientific community via the World Wide Web and by publication of a special issue of the journal *Proteins: Structure, Function, and Bioinformatics.* New categories, namely prediction of function, domain boundaries, and disordered regions have recently been included, but we will not discuss their results here. The CASP experiment has been extremely successful; it has been repeated every two years since it was first inaugurated and there is no sign it is going to be discontinued in the near future (Figure 2.10).

It has several merits. First of all, it has raised the issue of objective evaluation of structure prediction methods prompting the development of the continuous automatic assessment methods already described and fostering the development of similar initiatives in other fields, for example prediction of protein–protein interaction, gene finding, and scientific literature mining. It has also been instrumental in the development of common formats and evaluation measures in the field. It does also have limitations, however. Predictors in CASP are not necessarily in an ideal position to produce the best models, because of the time limitation imposed by the experiment. Also, because the results are public and very visible, predictors might not try "risky" innovations.

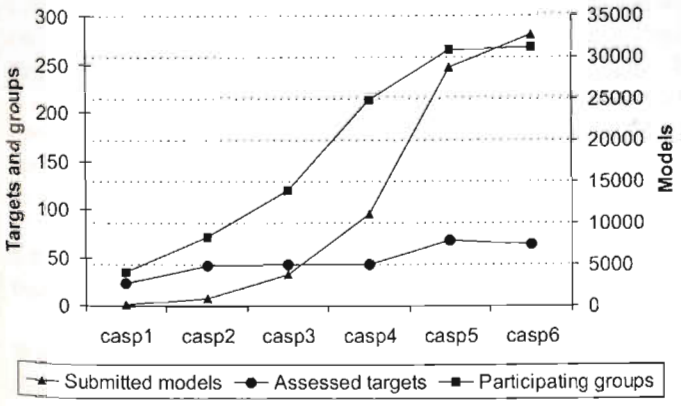


Figure 2.10 The plot shows the numbers of targets, participating groups, and models submitted to each of the editions of CASP from 1994 (CASP1) to 2004 (CASP6). All the thousands of models are publicly available on the CASP web site.

The assessors change every time, with rare exceptions, and are free to analyze the data in different ways and draw their own conclusions, although a set of measures is provided by the CASP organizers and is available for all the targets of all the editions. The most important are the GDT-TS values already described, *rmsd* values for several subsets of superimposed atoms, and the number of correctly aligned residues between target and model (a residue is considered correctly aligned if, after superposition of the experimental and modeled structure, its Cα atom falls within 3.8Å of the corresponding experimental atom, and there is no other Cα atom of the experimental structure that is nearer).

We will often refer to the CASP results in this book. We will also discuss how the CASP results can be used to measure progress between different editions, a non trivial issue, because it requires comparison of results obtained on a different set of targets, and at different times, therefore taking advantage of data bases of sequences and structures of very different size.

Suggested Reading

Description of most of the evaluation data are available on the CASP Web site (http://predictioncenter.org) where all the results of comparisons between models and target structures can be found.

The original definition of *SOV* has been published in:

B. Rost, C. Sander, R. Schneider (1994) Redefining the goals of protein secondary structure prediction. J Mol Biol 235, 13–26

A modification of the definition has been proposed in:

A. Zemla, C. Venclovas, K. Fidelis, B. Rost (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins, 34, 220–223

The web sites of EVA and Livebench are:
http://cubic.bioc.columbia.edu/eva/
http://bioinfo.pl/LiveBench/

The special issues of the journal "Proteins: Structure, Function and Bioinformatics" describing the results of the CASP experiment are published every two years by Wiley–Liss.

3
Ab-initio Methods for Prediction of Protein Structures

3.1
The Energy of a Protein Configuration

The native protein structure is the lowest free energy conformation that can be achieved kinetically by the polypeptide. It therefore seems natural to face the problem of predicting the structure of a protein by computing the free energy of every possible conformation and selecting the structure corresponding to the global free energy. The problem can be divided into two sub-problems – evaluation of the free energy of a given conformation and the search strategy for finding all possible conformations. Because the number of possible conformations of a protein is, as already mentioned, enormous, exhaustive search strategies cannot be employed; one must, instead, perform an approximate search strategy, the better the sampling of lower energy conformations, the better the search method.

In the next few pages we must discuss and understand some basic aspects of protein physics and, believe it or not, it is easier to do so by looking at equations. It is not always necessary to understand all the details of an equation – for example capturing the meaning of the Schrödinger equation requires a profound understanding of quantum-mechanics, and this is certainly not required to predict a protein structure. Hopefully, however, the reader will be able, by looking at the equations, to understand which variables are involved and the difficulty of the computation involved, without being distracted, or troubled, by the details of the formulas.

3.2
Interactions and Energies

When protein conformational energy is discussed, people talk almost indifferently about forces and energies. Although this can be confusing, remember that the two entities are directly related – the force is the derivative of the energy.

The behavior of a molecule can be completely described by the Schrödinger equation, the quantum mechanical equivalent of Newton's laws and of the law of conservation of energy in classical mechanics. The equation states that:

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

$$i\hbar\frac{\partial\Psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial\Psi}{\partial x^2} + V(x)\Psi(x,t) \quad (1)$$

where i is the imaginary unit, $\hbar$ is a constant, m is the mass of the particle, V is the potential, and Ψ the wave function. A wave function is a scalar function that describes the properties of waves – the Schrödinger equation predicts the future behavior of a dynamic system in terms of the probability of future events. What is important to understand is that to describe a system in terms of its energy we must consider all nuclei and electrons and their interactions. Even if we take into account the great difference between the masses of the electrons and nuclei, and treat them separately (Born-Oppenheimer approximation), it is impossible to solve the Schrödinger equation for systems larger than a few atoms. In practice, the equation is of relevance to protein structure prediction only insofar it can be solved for reasonably small systems and results from these model systems provide data for more approximate energetic calculations.

If we could solve the Schrödinger equation for reasonably large systems such as a protein and a set of surrounding water molecules, we could at least be certain that the energy of each conformation is computed correctly and we would be left "only" with the problem of exploring the conformational space of the system. In practice this is impossible. We need to find a function that approximately describes the energy of the interactions that occur in a protein molecule using a simplified representation of both the system (with the atoms being represented by points located at the center of their nuclei) and of the energetic contributions of each interaction in the protein. These interactions can, by and large, be divided into two types – covalent and non-bonded.

3.3 Covalent Interactions

A covalent bond between two atoms is formed if they share electrons (one pair in single bonds and two pairs in double bonds), but the effect is not localized and the electron density increase has an effect on the rest of the molecule. The standard way of approximating the potential energy for a bond is to treat the bond as a spring between the two atoms and describe its energy by use of Hooke's law (Figure 3.1):

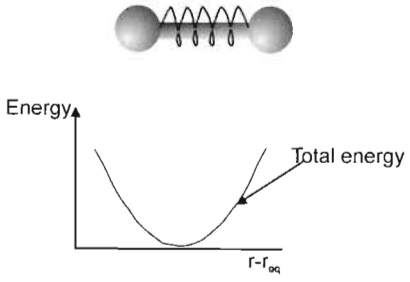


Figure 3.1 The classical approximation for a chemical bond.

$$E_{bond} = K_r(r-r_{eq})^2 \quad (2)$$

where r is the length of the bond, r_{eq} the equilibrium bond length, and K_r the spring constant (which is higher the stronger the bond).

Use of this equation is justified by the observation that bonds between chemically similar atoms in a wide variety of molecules have similar lengths (e.g. bonds from carbon atoms to hydrogen atoms are approximately 1 Å) so we can assume that the observed equilibrium value corresponds to the minimum potential energy.

Any function around its minimum value can be expanded as a Taylor power series:

$$f(x) = f(x_0) + \frac{\partial f(x_0)}{\partial x}(x-x_0) + \frac{1}{2}\frac{\partial^2 f(x_0)}{\partial x^2}(x-x_0)^2 + \frac{1}{6}\frac{\partial^3 f(x_0)}{\partial x^3}(x-x_0)^3 + \ldots \quad (3)$$

If we deform a bond around its equilibrium value, i.e. its energy minimum, the first derivative will be zero. Near the equilibrium, the expansion will be dominated by the term in x^2, therefore a function that varies around its minimum can be approximated by a quadratic function, $f(x) \approx A + Bx^2$. Because we are not interested in absolute values, the constant A can be neglected and this brings about the Hooke's law.

To approximate the energy needed to stretch a bond around its equilibrium value, we need to know the equilibrium length and the spring constant. The first is usually derived by analysis of small molecule X-ray crystal structures, the second by optimization or from quantum calculations on model systems. More accurate approximations can occasionally be used, but Hooke's law is usually used in most energy-evaluation procedures in protein structure prediction. It should be clear that this approximation can be used to compute the energy difference between length distances around their equilibrium value and has nothing to do with the energy of formation or breaking of a bond, in which case a quantum mechanical treatment is required.

Similar approaches can be used approximate the energy difference of variation of bond angles (Figure 3.2):

$$E_{bond} = K_\theta(\theta-\theta_{eq})^2 \quad (4)$$

For dihedral angles matters are slightly more complex, because they do not have a single energy minimum – indeed, they are usually represented by (Figure 3.3):

$$E_{dihedral} = \sum_{n=1}^{N} K_\phi[1+\cos(n\phi-\gamma)] \quad (5)$$

where N is the number of minima and γ is the angular offset.

In practice, it is found that this potential is not sufficient to represent the energy of a dihedral angle and often a non-bonded energy interaction term between the first and last atoms of the quadruplet is combined with Eq. (5).

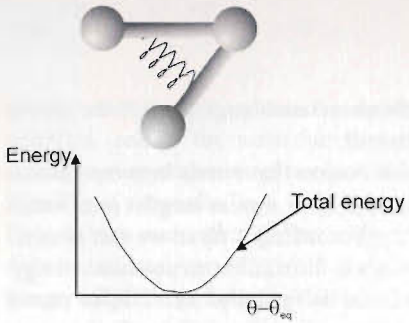


Figure 3.2 The classical approximation for the angle between three bonded atoms.

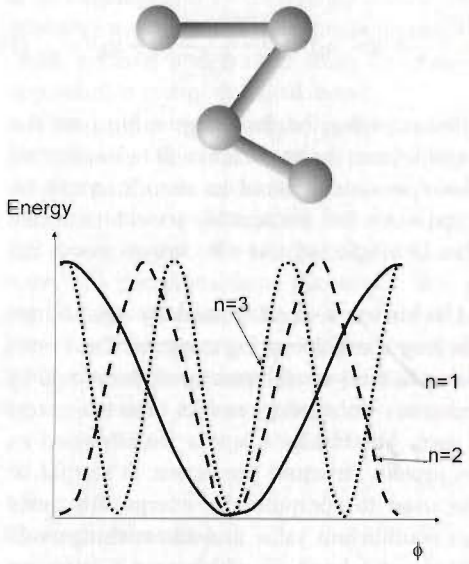


Figure 3.3 The approximation for a dihedral angle.

3.4 Electrostatic Interactions

The molecular scale is dominated by electromagnetic interactions. A nucleus and its electrons interact according to Coulomb's law (Figure 3.4):

$$E = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_r r_{ij}} \tag{6}$$

where q_i and q_j are the charges, r_{ij} their distance, ε_0 the permittivity, and ε_r the dielectric constant of the medium.

As we said, we cannot solve the Schrödinger equation to find the positions of nuclei and electrons, so we usually assign a "formal" charge to an atom, without explicitly considering its nucleus and its electrons. Some amino acids at physiological pH are neutral, but some are not and, furthermore, many proteins bind metal ions and the Coulombic interactions must be taken into account. As already mentioned, pairs of residues with opposite charges can form salt bridges that are strong interactions. These interactions are quite rare, especially in the core of proteins, because of the need for electrostatic interaction to compensate the solvation energy, i.e. the energy required to transfer a charge from a polar to a nonpolar solvent. Groups which carry no formal electrical charge can still be polarized, i.e. the orbitals can be distributed in such a way that parts of the molecule carry a charge. Some atoms have a tendency to attract electrons, and are, therefore, electronegative, whereas others have a tendency to lose electrons. The classical example of this is the water molecule, in which the electronegative oxygen attracts the electron and leaves the hydrogen atoms with net positive charges. Two water molecules can therefore form a strong electrostatic interaction, the hydrogen-bond, a bond approximately 2.8Å long and weaker than a covalent bond. Hydrogen bonds, as already discussed, are extremely important in protein structure. The tendency of main-chain carbonyl oxygen bonds to form hydrogen-bonds with main-chain amino groups leads to the possibility of forming different secondary structures; many side-chain groups can also form hydrogen-bonds.

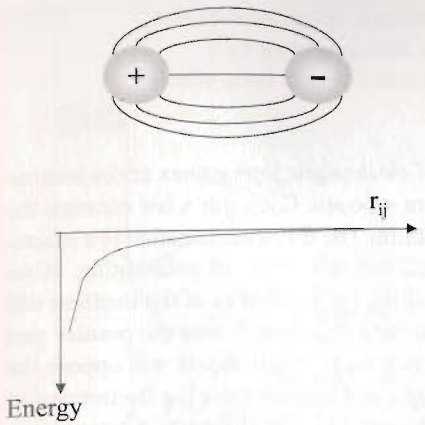


Figure 3.4 The Coulomb interaction.

Computationally, treatment of electrostatic interactions requires that an appropriate charge is placed at the position of the nucleus. These charges are partial to take into account the various effects neglected when using a classical approximation and their interaction is computed using the Coulomb's law.

Question: How do we compute the partial charge of an atom?

»The specific charges for each atom are computed by performing quantum mechanical calculations on model systems.

This approach is clearly a very crude approximation, and, indeed, calculation of electrostatic interactions is one of the

weakest points in the classical treatment of macromolecular interactions. Obviously the electric field generated by a charged atom polarizes other atoms and this, in turn, affects the charge of the atom.«

A further complication in the treatment of electrostatic interactions arises because proteins are not in a vacuum. As already mentioned, Coulomb's law contains the term ε_r, the dielectric constant of the medium. The dielectric constant is a macroscopic entity derived from the average microscopic effect of polarization. If we place two opposite charges in a polar medium, the molecules of the medium will tend to line up with the electric field with their dipole such that the positive part points towards the negative charge and vice versa. Their dipole will oppose the electric field, effectively reducing its strength, and thereby reducing the interaction energy between the two charged atoms (Figure 3.5). The dielectric constant takes this effect into account, in the hypothesis that the space around the charges is uniformly filled with a large number of molecules of the medium, but when we deal with proteins the distances between charges is of the same order of magnitude

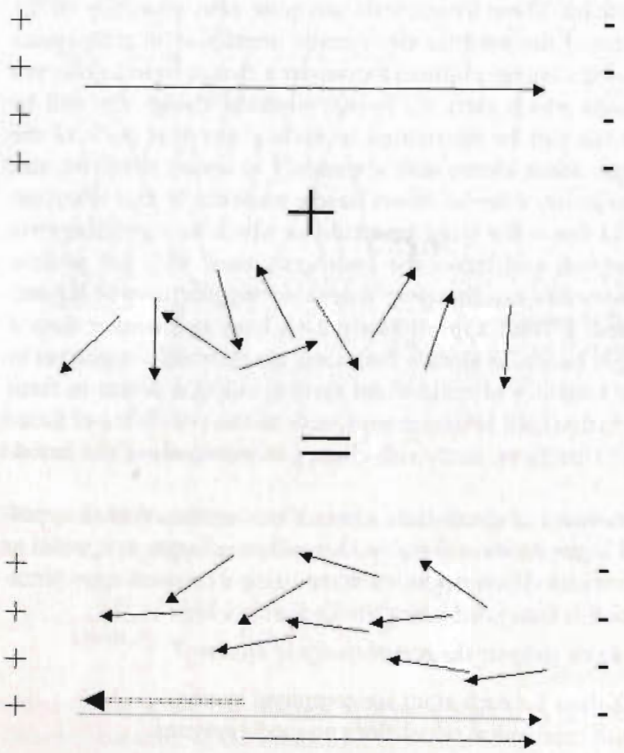


Figure 3.5 Polar molecules in an electric field orient themselves. The net effect is reduction of the electric field that is taken into account, macroscopically, by the dielectric constant.

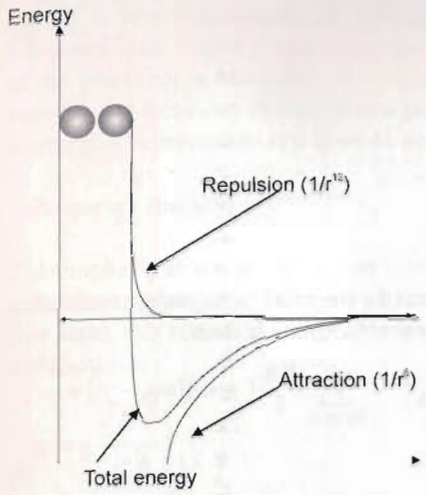


Figure 3.6 The van der Waals radius originates from interplay between the attractive and repulsive forces between two atoms.

as the size of the microscopic dipoles. The uniform distribution hypothesis breaks down and it is no longer possible to simply scale down the interaction by a fixed factor that takes into account the polarizability of the medium.

Sometimes, especially in the past, when computing power was limited, energy calculations were performed using a dielectric constant varying with the distance between the two charges (to take into account the fact that the farther away are two charges the more dipoles are between them), but this is a very crude approximation. Nowadays, it is more common to explicitly include in the calculation a large number of molecules of the solvent and use a dielectric constant of 1 or 2 (to take into account the electronic polarizability).

Electromagnetic interactions also affect uncharged atoms – they vibrate and this produces a dipole moment (because at any given instant the nucleus will be off center relative to the cloud of electrons) that interacts with the similarly generated dipoles of surrounding atoms. This produces an attracting interaction that has been shown to be in accordance with Eq. (7) (Figure 3.6):

$$E_{\text{dispersion}} = \frac{-B_{ij}}{r_{ij}^{6}} \tag{7}$$

where B_{ij} depends on the pair of atoms involved and is usually estimated empirically from data derived from small-molecule X-ray crystal structures.

The other effect that must be taken into account is that the orbitals of the atoms cannot overlap because of the Pauli exclusion principle that states that two electrons cannot have the same quantum state. The consequence is that two atoms cannot come too close to each other. This effect can be approximated by assuming that each atom is a hard sphere with a specific radius (the van der Waals radius) and two atoms cannot come closer than the sum of their radii. More realistically, although still approximately, this energy term is modeled as (Figure 3.6):

$$E_{\text{repulsion}} = \frac{A_{ij}}{r_{ij}^{12}} \tag{8}$$

where A_{ij} is an empirically derived term.

3.5 Potential-energy Functions

If we now take into account all we said about the interactions in protein molecules, we can write the potential energy of a given conformation C as:

$$E_C = \sum_{\text{bonds}} K_b(b_C - b_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta_C - \theta_{\text{eq}})^2 + \sum_{\text{dihedrals}} \frac{K\phi}{2}[1 + \cos(n\phi_C - \gamma)] + \sum_{\substack{\text{nonbonded} \\ \text{atoms}}} \left[\frac{A_{ij}}{r_{ij,C}^{12}} - \frac{B_{ij}}{r_{ij,C}^{6}} + \frac{q_i q_j}{\varepsilon_0 r_{ij,C}}\right] \tag{9}$$

This function gives the approximate value of the potential energy of a protein conformation. It does not include kinetics contributions to the energy and, to compute it, we must estimate a large number of terms such as the equilibrium values, the spring constants, etc.

Question: What Is the effect of using this approximate equation for the energy of a protein?

»The equation can be used to compute the approximate difference between two different conformations but gives no information about the free energy of forming that conformation and cannot be used to simulate quantum mechanical events such as chemical reactions, bond formation or breaking, etc.«

3.6 Statistical-mechanics Potentials

The atom-based approach to evaluation of protein potential energy is approximate and it has been shown on many occasions, including in the CASP experiments, that its energy terms are not sufficiently accurate for distinguishing the native conformation of a protein from an ensemble of reasonable alternatives. It is useful for quickly exploring some regions around the native conformation of a protein, or to evaluate whether a conformation is physically reasonable, but it still requires a fair amount of computation. An even more approximate, but nevertheless rather useful, approach is a residue-based strategy in which all atomic interactions between residues are attributed to a single point within each residue. This approach is knowledge-based, i.e. the energy contribution of each interaction is estimated from statistical analysis of the known structures of proteins. Calculation of the potentials is based on the heuristic observation that there is a correlation between the frequency of observing a particular structural feature, for example an interaction between two amino acids, and its energy:

$$\text{Frequency (feature)} \propto e^{-\beta E(\text{feature})} \tag{10}$$

This implies that we can count how many times we observe a given feature in the collection of proteins of known structure and, from this, compute its energy. At first sight, this seems to be nothing other than an application of the Boltzmann distribution:

$$p_i = \frac{e^{-\frac{E_i}{KT}}}{\sum_i e^{-\frac{E_i}{KT}}} \tag{11}$$

which states that, in a system at equilibrium, the probability of observing a state i is related to its energy E_i. In the equation, K is the Boltzmann constant and T the absolute temperature. Boltzmann's law is only applicable to an ensemble of identical but distinguishable particles in thermodynamic equilibrium. A database of protein structures consists of many different particles and it certainly cannot be regarded as a system at equilibrium. One way to try and understand why the equation is valid anyway is to think that if an interaction between two amino acids is energetically favorable there will be many protein sequences that can form that interaction. If we observe an interaction very frequently (i.e. in many different proteins with different sequences and structures), we can assume that the interaction is energetically favorable.

Knowledge-based potentials are derived by statistical analysis of known protein structures by counting how many times an interaction occurs. We can, for example, count how many times we observe a pair, for example, alanine–valine, at a distance of 3, 4, and 5 Å from each other. In practice we must introduce another term, their distance along the sequence. This is because we cannot expect our two amino acids to be at a distance incompatible with the number of peptide bonds connecting them. Our potential will have the form:

$$E(\text{a,b,d}) = -KT \ln p(\text{a,b,d})/p_0(\text{a,b,d}) \tag{12}$$

where p and p_0 are the observed fractions of total contacts observed between the two amino acid types a and b (for example alanine and valine), at a distance d (for example 5 Å) from each other in the known protein structures and in a reference state, respectively. In other words, p_0(a,b,d) represents the fraction of contacts that we expect to observe by chance alone. We need to count the occurrences of all possible pairs of residues, for different through-space and through-bond separations, and normalize these values in respect of a random ensemble of amino acids

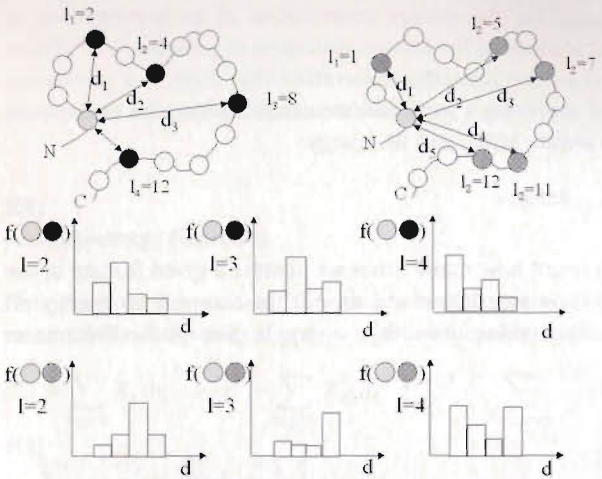


Figure 3.7 Schematic illustration of the method for deriving pair potentials. The number of contacts between the amino acid indicated by the black circle and that represented by the gray circle is counted for each of their separation distances in the sequences (the *l* values in the figure). This is repeated for each structure in the database and for each pair of residues and the frequencies are tabulated as a function of the type of amino acid, the separation distance, and the through-space distance. These values must be normalized by dividing them by the corresponding values observed for the background distribution of random structural arrangements.

(Figure 3.7). As often occurs in computational biology, defining the expected random distribution is the most difficult aspect, because we have no experimental information about the unstructured conformation of our protein.

We can generate many conformations with the same sequence as our target protein and count the observed interactions in this random ensemble, or we can reshuffle our sequence in the same target structures several times and count how many times the observed interactions arise by chance.

Question: How much does this choice affect the result?

»This choice is not without effect. If we use a set of random conformations, they will not, in general, show any secondary structure. Real proteins have a reasonable fraction of amino acids in secondary structures and the interactions corresponding to their geometry will show up more frequently than expected. For example, in alpha helices, residues separated by four bonds are at a fixed distance so that, if we compare a real protein structure distribution with that of a random polypeptide, we will observe a frequency peak at the corresponding distance for residues separated by three intervening amino acids. If, instead, we use real protein structures with reshuffled sequences as our reference distribution, they will also contain secondary structure. The helical pair-wise interactions in our protein structure will be more frequent than expected only for those amino acids that are in a helix more often than for a random sequence.«

When our potential have been derived, we can use it to evaluate the energy of a given conformation by adding the contributions of the observed interactions.

3.7 Energy Minimization

We have discussed how we can evaluate, at least approximately, the energy of a given protein conformation. We will discuss later in this chapter the limitations of our approximations. First, however, we will review the methods that are at our disposal to explore the conformational space of a protein, i.e. to modify the starting structure to obtain a structure with lower computed energy. Let us imagine we can explore every possible conformation of a protein and compute its energy using a potential energy function or a knowledge-based potential. A graph of the energy as a function of the conformation represents the energy landscape of the protein (Figure 3.8) and, if our computed energy were exact, the native structure would correspond to the conformation for which the energy has a global minimum:

$$E(\text{native conformation}) \leq E(\text{conformation}) \text{ for all conformations}$$

The problem of finding the global minimum of the energy function is difficult and, in general, no method can guarantee to find it, but there are techniques able to search for the local minima, i.e. for minima within a certain range from the

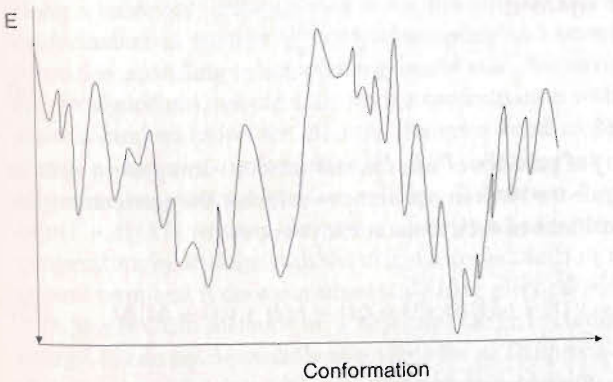


Figure 3.8 Each conformation of a protein has an associated energy, The graph shows a hypothetical plot of the energy as a function of protein conformation. In reality, each conformation is defined by a large number of coordinates, not just one as suggested by the figure.

starting value of the data (the initial conformation in our case). If we have a model of our protein we believe to be rather close in structure to the native conformation, energy minimization procedures could be used to vary the coordinates of the atoms within a certain range and find if any explored new arrangement has an energy lower than that of the starting structure. The approximations introduced in our treatment of the energy are too crude to guarantee that the computed energy difference between two similar conformations reflects their true energy difference with sufficient accuracy, therefore energy minimization is usually only used to remove unfavorable contacts or regularize hydrogen-bond lengths in a model, but there is no guarantee that our energy-minimized structure is closer to the native structure than our starting conformation.

3.8 Molecular Dynamics

Each atom of a protein has a potential energy and therefore feels a force exerted on it equal to the spatial derivative of this potential energy. We can compute the force acting on each atom and, within our classical approximation of a protein, simulate its motion by integrating the Newton' s second law of motion:

$$F_i = m_i \frac{\partial^2 x}{\partial t^2} \tag{13}$$

We can calculate the positions of each atom along a series of extremely small time steps and the resulting series of snapshots of structures over time is called a trajectory. What we do in practice is to select a temperature T and compute an initial atom velocity distribution that conforms with the total kinetic energy of the system according to the equation:

$$f = \left(\frac{m}{2\pi KT}\right)^{2/3} e^{-\frac{mv^2}{2KT}} \tag{14}$$

which gives the fraction f of particles of mass m and velocity v in a system with its temperature T. To integrate the Newton equation, we calculate the acceleration, the velocity, and the new positions of each atom at each time step:

$$a_i(t) = \frac{F_i}{m_i}; v_i(t + \Delta t) = v_i(t) + a_i(t)\Delta t; r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t)\Delta t \tag{15}$$

The time step must be sufficiently small to guarantee that the acceleration (i.e. the force) is practically constant during the step. An adequate time step for a molecular dynamics simulation is of the order of one femtosecond (10^{-15} s). The approximate time scales of the different motions in proteins are listed in Table 3.1. A simulation of a protein should be long enough to sample the motion several times and

therefore the length of our simulation must be directly related to the time scale of the phenomenon we wish to simulate.

Table 3.1 Time scale of different types of motion in a protein structure.

Motion	Femtoseconds (or number of steps in a typical molecular dynamics simulation)	Time (s)
Bond stretching	10	10^{-14}
Angle bending	100	10^{-13}
Rotating CH_3 groups	1000	10^{-12}
Water tumbling	20000	2×10^{-11}
Chemical reaction	1000000000	10^{-6}

Molecular dynamics simulations at relatively high temperature can be used to explore larger fractions of the protein conformational space while avoiding energetically "unreasonable" conformations, and they are, indeed, often used with this purpose by collecting snapshots of the simulation at regular time intervals and minimizing their energy.

3.9 Other Search Methods: Monte Carlo and Genetic Algorithms

Molecular dynamics samples the conformational space of a protein by moving along a trajectory, although this is not the only option. We can move from one conformation to another by random sampling and accept or reject the new conformation according to some energy-based rule. For example (Monte Carlo Metropolis algorithm), we can start from a conformation with potential energy E and make a random move, for example change a bond angle. We now compute the energy of the new conformation E'. If E' is lower than E, we accept the move, otherwise we extract a random number R between 0 and 1. If R is less than $\exp((E' - E)/KT)$ we keep the new conformation, otherwise we reject it. T is called temperature by analogy with the original applications of the method, but is just a control term and it does not necessarily have physical meaning.

In Monte Carlo methods we accept any change in conformation that reduces the energy, but do not necessarily reject the others. The probability of accepting a move that increases the potential energy depends on how much higher the potential energy of the new conformation is compared with that of the previous conformation, and on the selected temperature of the system. In this way we can move through the conformation space of the protein also exploring regions where the energy is higher, but not "too much higher", than that of our starting conformation (Figure 3.9).

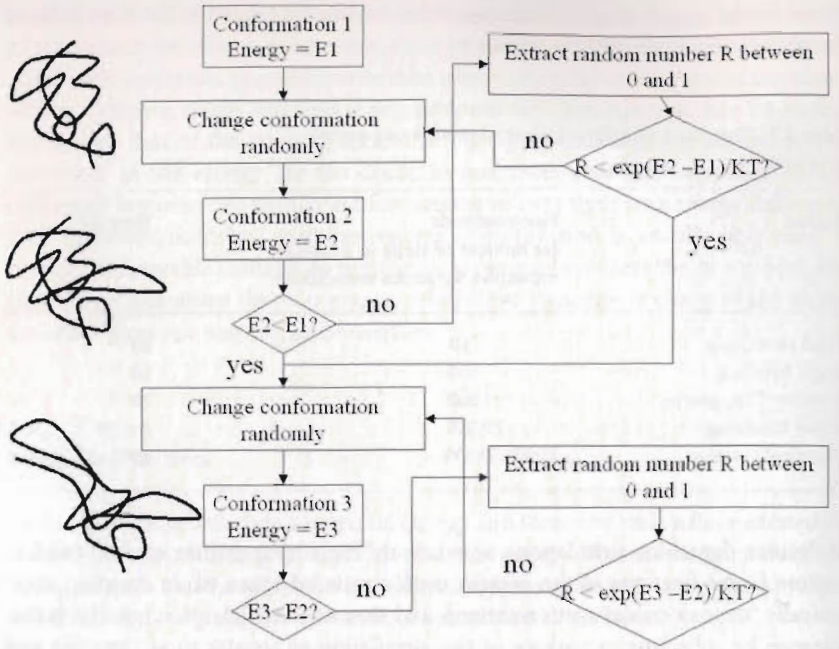


Figure 3.9 Schematic diagram of a simple Monte Carlo simulation.

Physically, our strategy is meant to ensure that the distribution of states at the beginning and at the end of our simulation have the same Boltzmann distribution, i.e. that the probability $p(i)$ of observing the conformation i of energy $E(i)$ is given by the equation:

$$p(i) \propto e^{\frac{-E(i)}{KT}} \tag{16}$$

The "temperature" of the simulation can be reduced after a number of steps, "cooling" the system. This process, called simulated annealing, can be thought of as an adiabatic approach to the lowest energy state.

Another search strategy is based on genetic algorithms that mirrors what nature does during evolution. In this algorithm linear strings of letters (representing the genome) are allowed to mutate, crossover, and reproduce. A genetic algorithm requires the definition of the initial population and of a fitness function. The initial population is usually encoded as a string of bits and the simulation starts by randomly applying "operations" to each individual. For example, a mutation involves exchanging the value of a bit (from 0 to 1 or from 1 to 0), a variation means incrementing or decrementing its bits by a small value (for example the string 0111 represents the number 7, by subtracting 1 from its value we obtain 6 which is encoded as 0110), a crossover means exchanging parts of one individual with parts of another. The next step involves ranking the individuals according to

their fitness and, for example, selecting the N with the highest fitness for a new round of simulation. The process is repeated until the desired distribution of fitness is reached, or after a predetermined number of steps. After several generations the population will consist of individuals that are well adapted in terms of the selected fitness function (Figure 3.10). The method does not guarantee that the final population contains the optimum solution for the given fitness function, but the procedure is more efficient than a random search.

Genetic algorithms are often used in protein structure prediction to generate a small set of native like conformations. The initial population is formed by several conformations of a protein and the fitness function can be the potential or knowledge-based energy, but we need a formalism to represent protein structures. We can make the genetic algorithm work on numbers rather than bits and encode our protein using the coordinates of the atoms, but in this case the probability that a mutation originates a chemically impossible protein structure is very high. More common is to encode the protein as a list of numbers representing the ϕ and ψ dihedral angles of its backbone, leaving bond distances and Ω angles fixed. In this representation the mutation operator requires replacement of one torsion angle with another (usually selected from among the values most frequently observed for the given residue), the variation operator will increment or decrement the angle by a preselected value, for example 5 or 10°, whereas crossover requires exchanging an N-terminal or a C-terminal, or an internal fragment between two conformations.

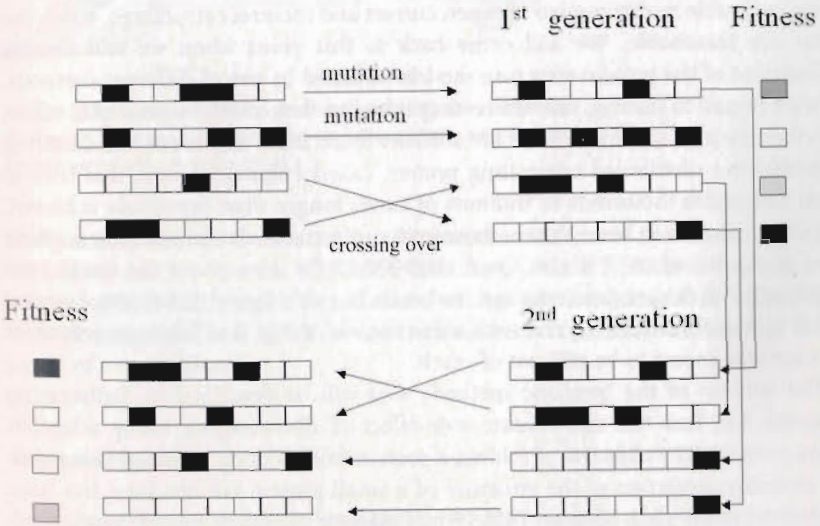


Figure 3.10 Simplified schematic diagram of a genetic algorithm. In the figure, the "genome", i.e. the encoding of the initial conformations, is made of two possible states, black and white. For a protein, this can be a bit string encoding its ϕ and ψ angles. The darker the box representing the fitness of each "individual", the higher its fitness. For proteins, this can be the value of the computed energy using, for example, a pair-wise potential. The next generation contains a number of copies of the "individuals" related to their fitness. For example, in the figure, the frequency of the fourth individual of the first generation is doubled in the second generation.

3.10
Effectiveness of Ab-initio Methods for Folding a Protein

Attempts to use molecular dynamics to fold a protein, i.e. starting from an unfolded conformation of a protein and letting it reach its global native structure, are countless. To simulate physical reality, a method must represent all the atoms of the protein and of the solvent, usually more than 10 000, use a small time step (approximately a femtosecond, as already remarked) and carry on the simulation for micro- or milliseconds. This is not yet feasible. Folding simulations have been used to study the unfolding processes of small proteins that can be accelerated by raising the simulation temperature and/or by applying external forces, and some encouraging results have been obtained. Care must still be taken when interpreting the results, however, because the short-time simulations can only sample a very limited conformational space.

It is not yet clear whether, even if one could run a very detailed simulation for a very long time, our treatment of the protein as a classical object is sufficiently accurate to be useful in folding simulations of larger, and therefore more complex, proteins. None of the available energy calculation methods and force fields yet seems able to consistently detect the lowest-energy conformation among a set of "decoy" conformations. For example the computed potential energy of a crystal structure is not necessarily found to be lower than that of related, but different, conformations of the same protein. In other words, the computed energy does not seem to be able to distinguish between correct and incorrect structures, when the latter are reasonable. We will come back to this point when we will discuss refinement of the protein structure models obtained by use of different methods.

With regard to folding, one interesting initiative that could, in principle, tell us whether the available force fields are sufficiently accurate to achieve this daunting objective is a distributed computing project, called Folding@home, that tries to span timescales thousands to millions of times longer than previously achieved. The idea consists in letting users download and run simulation software on their own computer when it is idle. Over 1 000 000 CPUs throughout the world have participated in the project in the last few years, but although simulations of several small proteins have already met with some success, the goal of folding a protein of average size seems to be still out of reach.

The success of the heuristic methods that will be described in forthcoming chapters has had the unfortunate side-effect of discouraging many scientists from pursuing the objective of folding a protein from scratch, although a successful ab-initio prediction of the structure of a small protein (48 residues) has been submitted to the sixth edition of CASP by Professor Scheraga, one of the pioneers of the field of molecular simulations (Figure 3.11). It is important that such efforts are not discontinued. We can assert that we have understood the protein-folding process only if we can reproduce it. Even if we could predict the structure of all the proteins of the universe by using the empirical knowledge-based methods that will be described in the rest of this book, we would still be lacking a deep understanding of the physics of the folding process. This is clearly unsatisfactory, both intellectually and otherwise. Understanding the process would have practically important consequences, because the principles that drive protein folding also dictate substrate and ligand binding and the induced conformational changes often associated with protein functions and explains the process of incorrect folding that is the cause of many diseases.

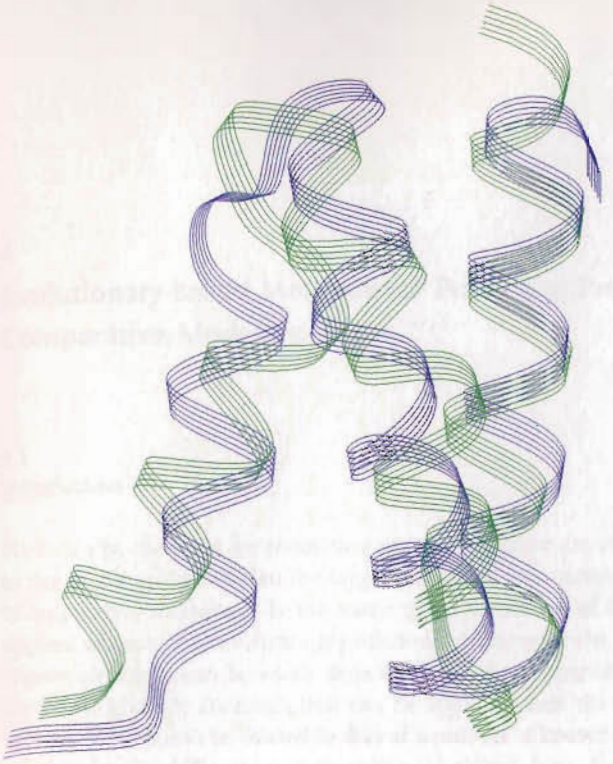


Figure 3.11 Comparison of a prediction submitted in CASP6 by Professor Scheraga (in green) and its subsequently determined experimental structure (in blue).

Suggested Reading

A seminal paper discussing the problem of modeling a protein:
C. Levinthal (1966) Molecular model-building by computer. Scientific American 214, 42–52

An extremely good and interesting book on proteins physics:
A.V. Finkelstein, O. B. Ptitsyn (2002) Protein Physics. Academic Press, London