Chapman & Hall/CRC Mathematical Biology and Medicine Series

The Ten Most Wanted Solutions in Protein Bioinformatics

Anna Tramontano

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York Singapore

Published in 2005 by
Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-4916 (Hardcover)
International Standard Book Number-13: 978-1-5848-8491-0 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Tramontano, Anna.
The ten most wanted solutions in protein bioinformatics / Anna Tramontano.
p. cm. -- (Chapman & Hall/CRC mathematical biology and medicine series)
Includes bibliographical references and index.
ISBN 1-58488-4916 (alk. paper)
1. Proteomics. 2. Bioinformatics. I. Title. II. Series.

QP551.T723 2005
572'.6--dc22 2005041404

Catalog record is available from the Library of Congress

T&F informa
Taylor & Francis Group
is the Academic Division of T&F Informa plc.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Foreword

The goal of protein bioinformatics is to assist experimental biology in assigning a function or suggesting functional hypotheses for all known proteins. The task is formidable. A simple calculation shows that we cannot possibly study each and every biological molecule of the universe. Therefore, we need fast and reliable computational methods to extrapolate the knowledge accumulated on a subset of cases to the rest of the protein universe.

This book reviews available methods in protein bioinformatics, with a special emphasis on their effectiveness in inferring the biological properties and functional roles of proteins. It is organized around specific problems that elicit the efforts of the community, and it focuses on the limitations of current approaches and on future developments that are likely to improve our understanding of the exquisitely specific and efficient mechanisms of protein function.

Bioinformatics is an interdisciplinary science that synergistically utilizes the contributions of informatics, physics, and mathematics, but, ultimately, the objective is the solution of biological problems. Therefore, this book starts with an overview of what we know about the structure and function of proteins. Proteins are a product of evolution. Thus, the basic principles of evolution must be kept in mind when new methods are devised or new routes are explored for inferring the function of a biological macromolecule. This book addresses the problem of detecting the existence of an evolutionary relationship between proteins in Problem 1. The detection of local similarities between protein sequences and the analysis of high-throughput experiments can also be effectively exploited for functional assignment, as is shown in Problems 2 and 3. Much more information can be derived from the knowledge of the three-dimensional structures of proteins. These structures can be experimentally determined or inferred from computational methods (Problems 4 and 5) and studied for insight into the roles of proteins (Problem 6). Proteins interact with each other and with ligands, both physically and logically (Problems 7 and 8), as parts of complex regulative networks. Several methods are being devised to explore these aspects of protein function. Finally, we discuss the extent to which our understanding of proteins allows us to design completely new proteins tailored to specific tasks (Problem 9) or to rationally modify the function and properties of existing proteins (Problem 10).

As we will see, many unsolved problems remain in each of these areas, and new ideas are continuously being produced and tested. The pressure on this relatively new discipline is strong because an understanding of life, in all its beauty and complexity, finally seems within our reach, and our astonishment at being so close to our goal is only equaled by our impatience to reach it.

Introduction

Proteins are the major components of living organisms and constitute more than 25% by weight of a typical cell. Even more impressive is the variety of functions that they can perform: catalysis, immune recognition, cell adhesion, signal transduction, sensory capabilities, transport, movement, and cellular organization. From a chemical perspective, proteins are linearly-oriented heteropolymers of amino acids (small organic molecules) whose structures and properties are described in this book. The sequence of amino acids in a protein is determined by the sequence of nucleotides, or bases, in the corresponding gene. Each adjacent triplet of bases of a gene in the DNA codes for one amino acid or for a codon that signals the end of the gene, according to the practically universal genetic code shown in Table 1.

The nucleotide sequence of a genomic region is technically much easier and faster to obtain than the sequence of the encoded protein, as is evidenced by the pace at which the complete genomes of many organisms, including *Homo sapiens*, are being deciphered. The large majority of known protein sequences are in fact deduced from the corresponding sequences of the genes, rather than from direct chemical sequencing of the proteins.

TABLE 1

The Genetic Code

First Base		**Second Base**					**Third Base**
		U	**C**	**A**	**G**		
	U	Phe (F)	Ser (S)	Tyr (Y)	Cys (C)	U	
		Phe (F)	Ser (S)	Tyr (Y)	Cys (C)	C	
		Leu (L)	Ser (S)	Stop	Stop	A	
		Leu (L)	Ser (S)	Stop	Trp (W)	G	
	C	Leu (L)	Pro (P)	His (H)	Arg (R)	U	
		Leu (L)	Pro (P)	His (H)	Arg (R)	C	
		Leu (L)	Pro (P)	Gln (Q)	Arg (R)	A	
		Leu (L)	Pro (P)	Gln (Q)	Arg (R)	G	
	A	Ile (I)	Thr (T)	Asn (N)	Ser (S)	U	
		Ile (I)	Thr (T)	Asn (N)	Ser (S)	C	
		Ile (I)	Thr (T)	His (H)	Arg (R)	A	
		Met (M)	Thr (T)	His (H)	Arg (R)	G	
	G	Val (V)	Ala (A)	Asp (D)	Gly (G)	U	
		Val (V)	Ala (A)	Asp (D)	Gly (G)	C	
		Val (V)	Ala (A)	Glu (E)	Gly (G)	A	
		Val (V)	Ala (A)	Glu (E)	Gly (G)	G	

Note: Each triplet of bases in a gene codes for one of the 20 amino acids, here listed in their three-letter and one-letter codes.

However, the genetic material not only contains genes, but also contains regulatory regions and noncoding regions of unknown function, such as long and short repeats; pseudogenes and retropseudogenes; satellite, minisatellite, and microsatellite regions; transposons and retrotransposons; viral vestigials; and others. In higher organisms, the gene sequence is also interrupted by noncoding fragments of variable length called introns.

Although the large body of available genetic information holds the promise of unraveling the meaning of life, we must decode this information; that is, we must detect which regions are the gene-coding regions, translate these regions into the corresponding protein sequence, and work out the protein's molecular function. The development of methods for finding the genes and their corresponding proteins and for unraveling their function is essential. It is the only route to utilizing our genomic knowledge for rationally interfering with diseases and understanding, for example, the genetic basis of individual pharmacological responses.

In this book, we do not discuss the problem of finding genes, which is a major challenge that the genomic era is posing to bioinformatics. Rather, we concentrate on the techniques that can be applied to derive functional knowledge of a protein, once the complete sequence of its amino acids is known.

The Structure of Proteins

The function of a protein depends upon its "shape;" that is, upon the three-dimensional structure that can be determined by X-ray crystallography or nuclear magnetic resonance experiments. The resulting data are stored in a data base called the Protein Data Bank (PDB). At present the PDB contains a few thousand examples of protein structures, but it is rather redundant. Often, different examples of the structure of a protein have the same amino acid sequence but in different states, such as with different bound ligands, in complex with different proteins, or determined under different experimental conditions. The database contains more than 800 entries for the protein lysozyme, for example.

A polymer does not necessarily assume a unique three-dimensional structure in solution, which is equivalent to saying that its energy landscape (the value of the free energy for each possible arrangement of its atoms) does not necessarily have a single, global free-energy minimum. However, a protein is not just any polymer. It is a special polymer in that, in a given environment and physiological conditions (pH, temperature, ionic strength, etc.), it assumes one, and only one, specific three-dimensional structure. Some important implications and limitations of this statement are discussed in Problem 4.

Figure 1 shows the experimentally determined three-dimensional structure of a protein in which each atom is depicted as a sphere (see color insert after page 40). The protein is glycogen phosphorylase, one of the enzymes that allow us to survive without feeding continuously, even though our cells need

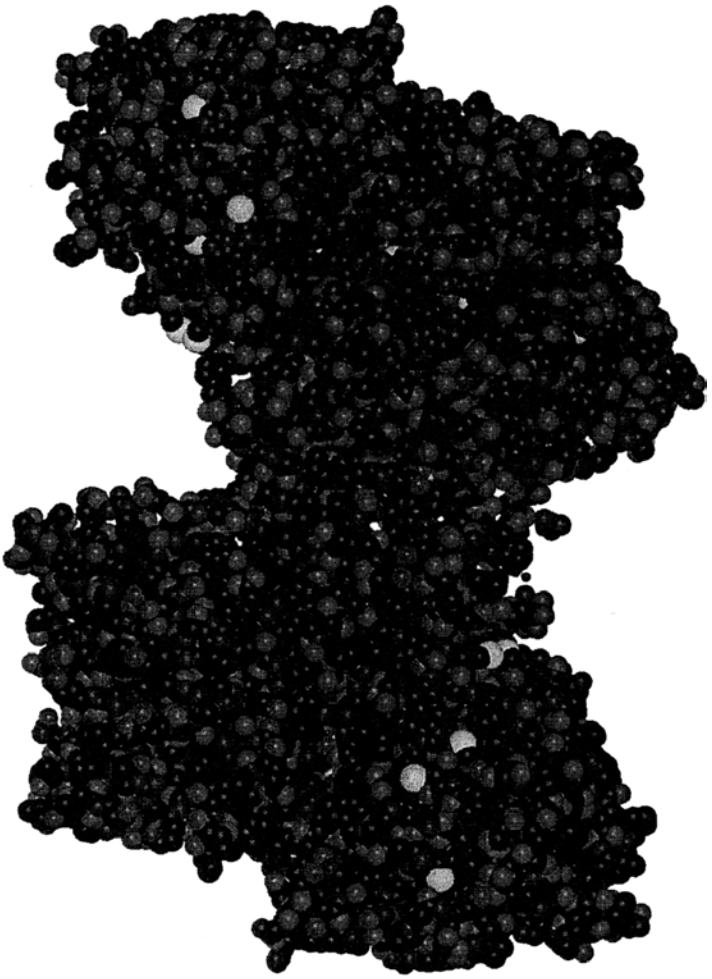

FIGURE 1
An all-atom representation of a protein structure determined by X-ray crystallography. This protein is an enzyme, glycogen phosphorylase from rabbit muscle, and its code in the Protein Data Bank is 1ABB. Atoms are colored according to a commonly used scheme: carbon is black, nitrogen is blue, oxygen is red, and sulfur is yellow.

a constant supply of sugars. The sugars that we consume are stored in our muscles in the form of glycogen, a macromolecule that contains up to 10,000 glucose molecules. The glycogen granule is clipped into glucose, and this chemical reaction is catalyzed by glycogen phosphorylase.

At first sight, the structure appears very complex, with no apparent regularities. Hopefully, though, by the end of this Introduction, the reader not only will be fascinated by the beauty of this molecule and impressed by its versatility, but also will have learned how to detect the underlying regularities as well as some general properties of protein structures.

FIGURE 2
A different representation of the protein in Figure 1. This time, each amino acid is depicted as a vertex of a broken line connecting the amino acid chain.

The remainder of this Introduction is devoted mainly to the structure of proteins that spend their time in polar environment. Apolar proteins, which are embedded in biological membranes, and their structural properties are discussed in Problem 5.

Figure 2 shows the same protein as in Figure 1, but now, instead of every atom, only one atom per amino acid is shown as a vertex of a broken line that connects equivalent atoms in consecutive amino acids. The atom selected in the figure is called Cα and, in amino acids, it is linked to four different chemical groups: a carboxylic group, an amidic group, a hydrogen atom, and a variable chemical group (the side chain). The amino acids that occur in natural proteins number exactly 20 and differ by their side chains (as shown in Figure 3). The side chain can be a single hydrogen atom, as in the case of glycine, or can contain polar, neutral, and charged groups. Hydrogen atoms are usually not shown in protein-structure representations, because their positions are difficult to detect by X-ray crystallography.

Amino acids are linked to each other by a chemical bond, the peptide bond, between the carboxylic group of one amino acid and the amidic group of the adjacent amino acid. The chemical chain formed by the amidic group, the Cα, and the carboxylic group is called the main chain, or backbone, of the polypeptide. Different side chains protrude from the backbone, and their

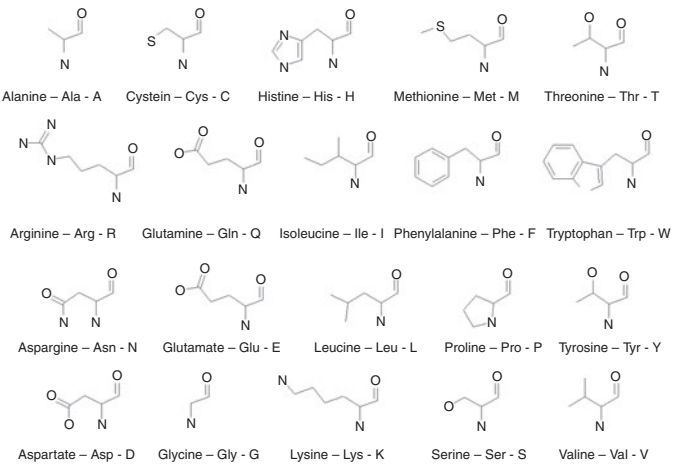

FIGURE 3
The 20 naturally occurring amino acids.

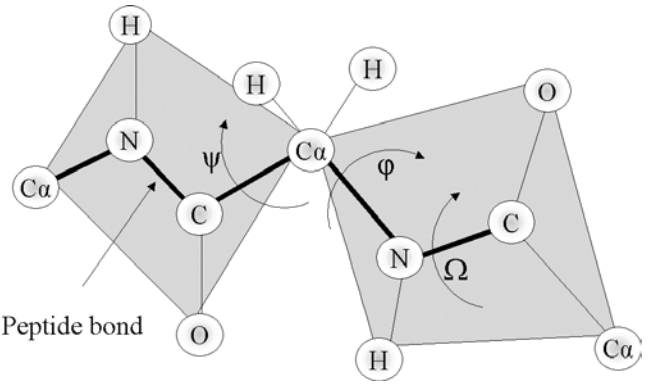

FIGURE 4
Amino acids linked by a peptide bond. The thicker lines form the "backbone" of a protein.

sequence defines the properties of the protein. The sequence of amino acids of a protein is called its primary structure.

The backbone of a polypeptide chain is quite flexible (as can be appreciated by looking at Figure 2). However, only the two angles ϕ and ψ (Figure 4) can assume several conformations in solution, the remaining angle, around the peptide bond, is planar. Furthermore, not all combinations of the values of ϕ and ψ are energetically favorable. Some are rarely observed, as is shown in Figure 5.

The combinations ϕ –60 and ψ –50 and ϕ –110 and ψ 130 are energetically favorable and observed very often. A consecutive stretch of residues with ϕ and ψ values in the first region, called the α_R region, assumes a helicoidal

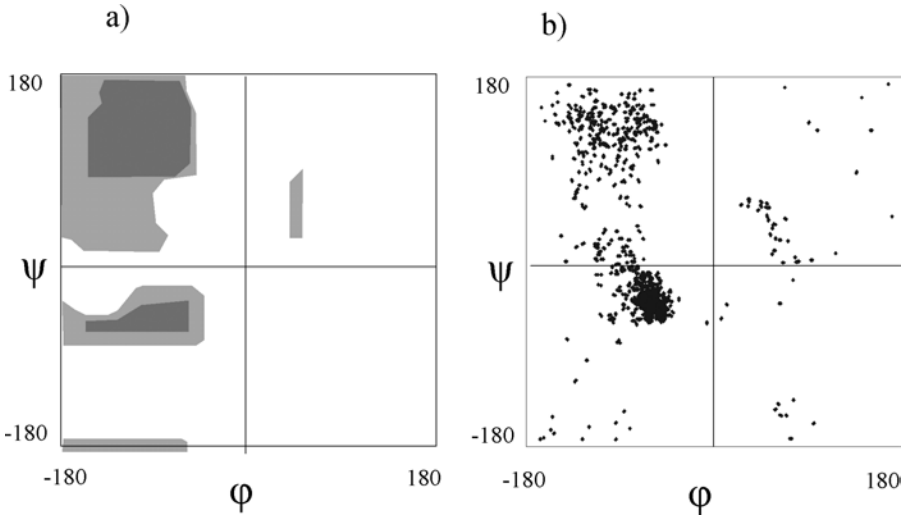

FIGURE 5
On the left is a representation of the results of energetic calculations of all possible pairs of ϕ and ψ angles in a dipeptide formed by alanine residues. Some combinations are energetically favorable (dark-gray areas) or allowed (light-gray areas), whereas others are unfavorable (white areas). This arrangement is reflected by the frequency at which combinations are observed in experimentally determined protein structures, as on the right side of the figure, where each point represents a ϕ and ψ pair observed in glycogen phosphorylase. The region with φψ angles around (60, 60) is rarely observed, and it is generally unfavorable because it brings the first carbon atom of the side chain too close to the carboxylic oxygen. The amino acid glycine does not contain a carbon in its side chain and is often observed in this conformation. The graphs shown in the figure are called Ramachandran plots.

shape. A stretch with values in the second region, the β region, becomes elongated and forms hydrogen bonds with other regions with the same local structure, as is shown in Figure 6 (see color insert after page 40).

Note that the polar atoms of the backbone (the carboxylic and amidic group) of both the α-helix and the β-sheet form hydrogen bonds with other main-chain atoms. This behavior is a result of the fact that, in the unfolded state and in an aqueous environment, the polar atoms would form hydrogen bonds with the surrounding water molecules. When the protein folds (i.e., when it assumes a compact shape), some of these atoms are shielded from the solvent and unable to form hydrogen bonds with it. This energy loss has to be compensated by the formation of hydrogen bonds within the protein chain.

A protein chain, in general, contains both hydrophobic and hydrophilic atoms. Exposure of the former to a polar solvent is energetically unfavorable, because a loss of entropy results (Figure 7). During folding, an energy gain is associated with the shielding of these groups from the solvent in addition to an energy gain through internal interactions (vanderWaals, charge–charge, and intrachain hydrogen bonds) established in the final structure. In proteins, these interactions are sufficient to compensate for the loss of entropy

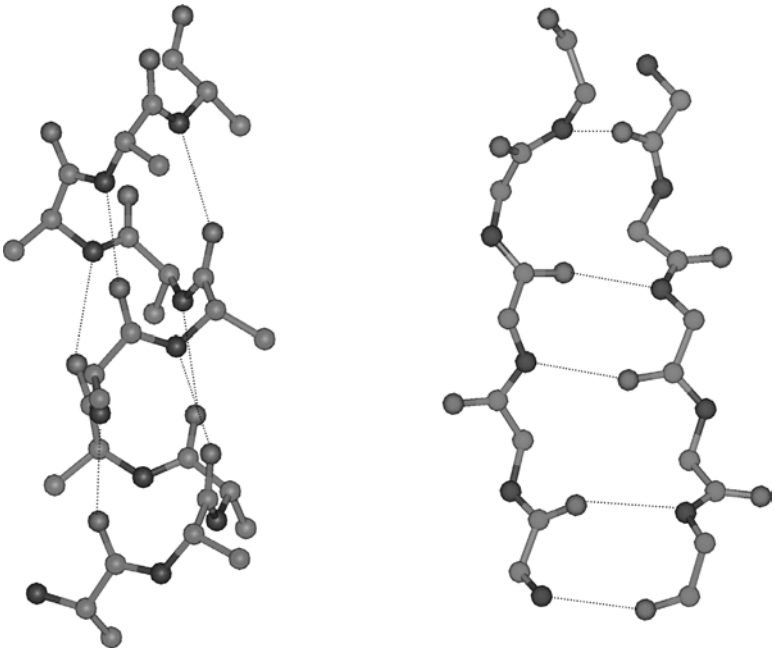

FIGURE 6
The backbone atoms of an α-helix and of two β-strands are depicted above. The strands, pairing via hydrogen bonds (dotted lines), form a β-sheet.

associated with folding (an unfolded chain has practically an infinite number of possible conformations and, therefore, a very high entropy), and a unique three-dimensional structure (called tertiary structure) can be achieved.

In a protein structure, polar amino acids (i.e., amino acids whose atoms can form hydrogen bonds with the water) are found more often at its surface, while the hydrophobic amino acids are mostly buried inside (Figure 8) (see color insert after page 40).

Most proteins contain regions in α and β conformations, collectively called regions of repetitive secondary structure, and connecting regions called loops. We can further modify our view of proteins by using cylinders and arrows to depict the secondary structure elements, as is shown in Figure 9 (see color insert after page 40). The latter representation of our protein shows its beautiful regularity. Two seemingly identical chains (a protein formed by more than one amino acid chain has a quaternary structure) are formed by two "lobes"; that is, structural regions that have more contacts between themselves than with other regions of the protein. We call these "lobes" domains. The two domains are not identical, but they show some topological similarity and can be described as three-layered structures, two external helical parts and a central β-sheet.

Cellular mechanisms can chemically modify a protein's primary structure after it has been synthesized. These modifications can be permanent or can

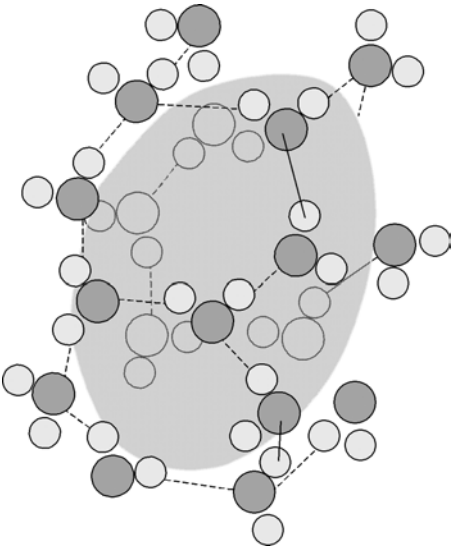

FIGURE 7
The hydrophobic effect. Polar molecules (water in the figure) form many energetically favored hydrogen bonds. When an hydrophobic molecule is present, they organize themselves around it in a more ordered way and therefore lose entropy.

vary according to the cellular state. For example, several proteins are glycosylated; that is, sugar chains are covalently linked to their amino acids. Another common protein modification is phosphorylation, which is often used for regulation, signal transduction, and cell cycle regulation. These modifications can affect the protein structure, often to a very considerable extent, and their presence and precise localization can depend on specific patterns of amino acids.

Glycogen phosphorylase is one such example. When a phosphate molecule is added to a serine amino acid (serine 14), shown in Figure 10 as a green sphere (see color insert after page 40), a shift occurs in the structural elements of the enzyme. This conformational change activates the protein. Phosphorylation of this enzyme is performed by other enzymes that monitor the concentration of sugar in the blood. The activity of the protein also has to increase when the energy levels of the cell are low. AMP (adenosine monophosphate) is a product of ATP (adenosine triphosphate) breakdown, an energetically favorable chemical reaction that provides energy to the cell. More AMP is created when energy levels are low and more sugar is needed. Binding of AMP to a site on glycogen phosphorylase causes similar structural changes as phosphorylation and activates the enzyme.

The Structure–Function Relationship in a Protein

The amino acid sequence of a protein contains amino acids selected for shaping its energy landscape that specify the unique three-dimensional native

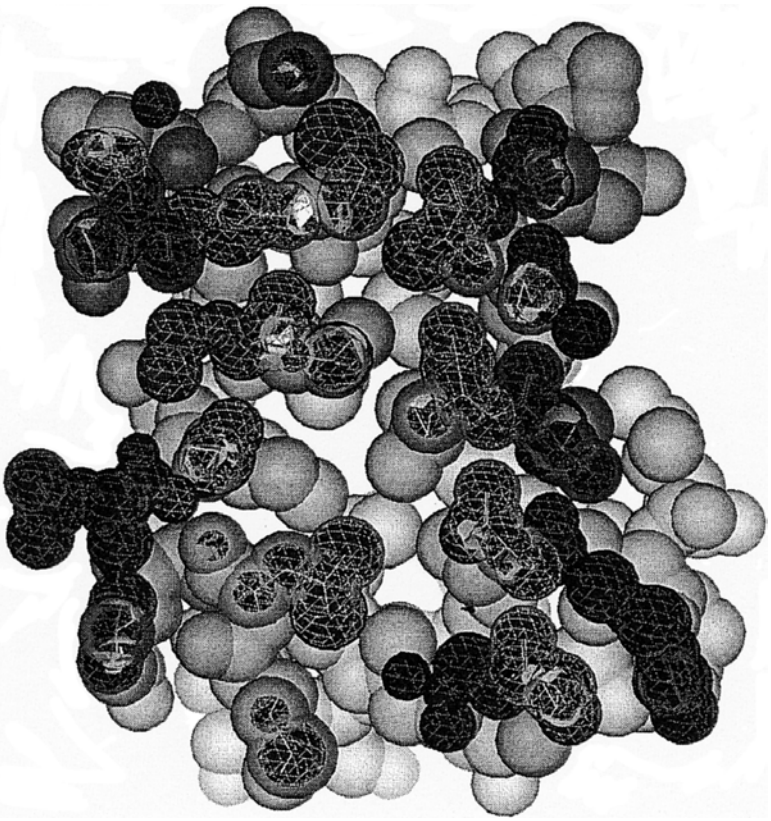

FIGURE 8
A section of the structure of SH3, a small module found in many proteins, where it acts as an adapter to recruit other proteins. The green hydrophobic amino acids are more frequent in the inside than on the outside of the molecule.

structure and do not allow chains to fold into undesired conformations or not achieve a definite structure at all. The amino acid sequence also contains the specific residues necessary for the protein's function.

Without delving into the chemical details of the enzymatic activity of glycogen phosphorylase, we mention that it needs the close proximity of four amino acids (see the inset in Figure 10): two lysine residues at positions 568 and 574, one arginine at position 569, and one threonine at position 676. These amino acids are distant in the linear amino acid chain, but are brought together in the precise relative position needed for catalysis by the three-dimensional structure of the protein in its active form (i.e., when the protein is phosphorylated or when AMP is bound). The region of an enzyme where action takes place is called its active site. The structure of the active site in the inactive form of glycogen phosphorylase (i.e., the relative position of the catalytic residues) has a somewhat different structure than in its active form. The enzyme is, in fact, less efficient. Incidentally, the existence of two structures

FIGURE 9
The structure of glycogen phosphorylase once again. This time helices and strands are shown as cylinders and arrows.

for the enzyme does not contradict what we said about the uniqueness of the protein shape, because the two conformations are achieved with different ligands and, therefore, are not in the same environmental conditions.

Another example of the importance of the three-dimensional structure for function is illustrated in Figure 11 (see color insert after page 40). The reader should now see that the depicted protein is mostly formed by β-strands. It has two domains and a quaternary structure. It is an enzyme encoded by the virus responsible for hepatitis C. This virus enters the host cell and synthesizes a single, long amino acid chain that is later broken into smaller fragments, each of which encodes one of its functions. The enzyme shown in the figure breaks up the long polyprotein. It is a protease; that is, an enzyme that catalyzes the cleavage of peptide bonds.

We can now look at its active site. The atoms responsible for the catalytic activity belong to three amino acids: a serine, a histidine, and an aspartic acid. It also has a "pocket" ideally suited for accommodating the side chain of one specific amino acid, cysteine, and, in this way, it recognizes the precise location where cleavage should occur. In the figure is an enlarged view of the region involved in recognition and catalysis. The involved amino acids come from different parts of the amino acid chain and, once again, the three-dimensional structure of the protein allows them to be correctly positioned.

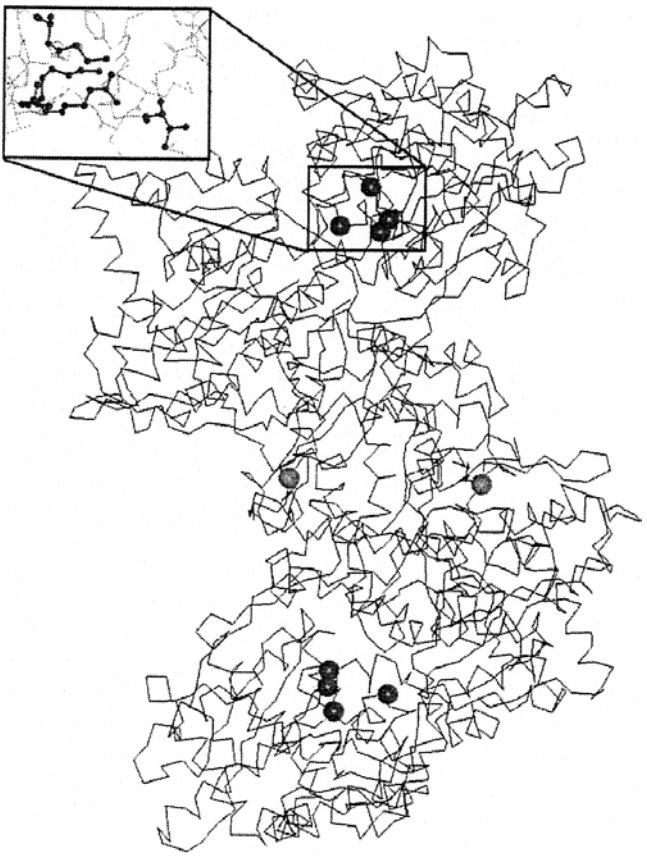

FIGURE 10
The active site of glycogen phosphorylase. The phosphorylation of serine 14, shown as a green ball, triggers a conformational change in the protein.

One way to inhibit the activity of this enzyme and, thereby, interfere with the function of the virus is to design a molecule that occupies and blocks the site where the enzyme binds the target amino acid (a cysteine). Detailed knowledge of the three-dimensional structure of the protein is fundamental to designing such a molecule.

Detection of the residues responsible for function (e.g., those that form an active site or an interaction surface) solely on the basis of a protein's amino acid sequence is practically impossible. These residues are no different from other amino acids. Only their specific positioning in the context of the final three-dimensional structure allows them to perform their function. Yet, the goal we are pursuing is detection of the sites important for activity and the understanding of how a function is performed, in the absence of an experimental three-dimensional structure.

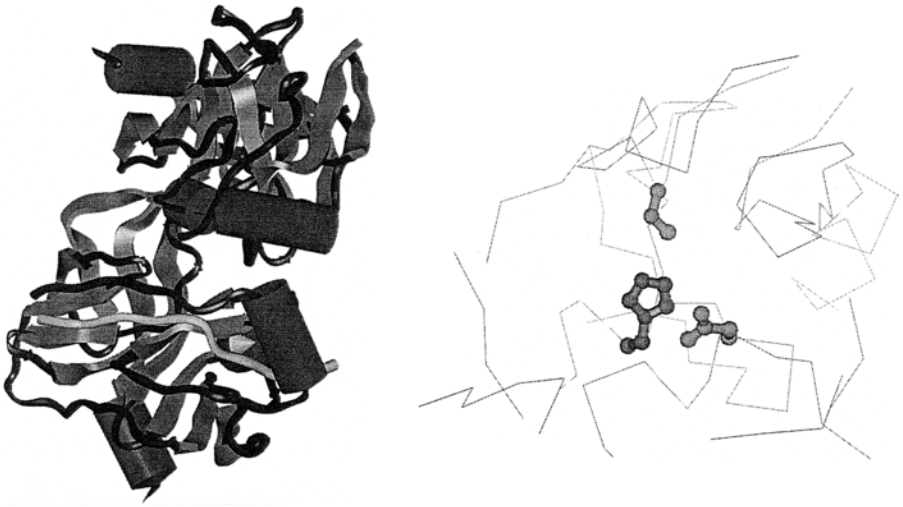

FIGURE 11
The structure of the protease of the hepatitis C virus (PDB code: 1NS3).

Suggested Reading

Berg, J.M., Stryer, L. and Tymoczko, J.L. *Biochemistry,* 5th ed., W. H. Freeman, New York, 2002.

Branden, C. and Tooze, J. *Introduction to Protein Structure*, 2nd ed., Garland Publishing, New York, 1999.

Creighton, T.E. *Proteins*, 2nd ed., W.H. Freeman, New York, 1993.

Nelson, D.L. and Cox, M.M. *Lehninger Principles of Biochemistry,* 4th ed., W. H. Freeman, New York, 2004.

Lesk, A.M. *Introduction to Protein Architecture—The Structural Biology of Proteins*, Oxford University Press, Oxford, 2000.

Voet, D. and Voet, J. *Biochemistry,* 3rd ed., Wiley, New York, 2004.

Drenth, J. *Principles of Protein X-ray Crystallography*, 2nd ed., Springer, New York, 1999.

Rhodes, G. *Crystallography Made Crystal Clear*, Academic Press, New York, 1999.

Wuthrich, K. *NMR of Proteins and Nucleic Acids*, Wiley-Interscience, New York, 1986.

Cavanagh, J., Arthur, W.J.F., Palmer, III, G., and Skelton, N.J. *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, New York, 1995.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures, *Eur. J. Biochem.* 80, 319–324, 1977.

Acknowledgments

Numerous people have generously offered me advice and support. I express my gratitude to all colleagues of the Department of Biochemical Sciences of the University of Rome "La Sapienza." In particular, I thank professors Maurizio Brunori, Francesca Cutruzzolà, and Carlo Travaglini-Allocatelli and doctors Veronica Morea, Romina Oliva, and Simonetta Soro. My most special thanks go to Dr. Domenico Cozzetto for his dedication and patience in critical reading of the manuscript.

Contents