METHODS FOR PROTEIN ANALYSIS

A Practical Guide to Laboratory Protocols

ROBERT A. COPELAND

CHAPMAN & HALL
New York • London

First published in 1994 by
Chapman and Hall
One Penn Plaza
New York, NY 10119

Published in Great Britain by
Chapman and Hall
2-6 Boundary Row
London SE1 8HN

Printed in the United States of America on acid free paper.

Library of Congress Cataloging in Publication Data

Copeland, Robert Allen.
Methods of protein analysis : a practical guide to laboratory protocols / Robert A. Copeland.
p. cm.
C. Costa, 212-244-3336, Chapman & Hall
Includes bibliographical references.
ISBN 0-412-03741-6
1. Proteins—Analysis—Laboratory manuals. I. Title.
QP551.C725 1993
547.7'5'028—dc20 93-6934
CIP

British Library Cataloguing in Publication Data also available.

Preface

With the revolution that has occurred in Molecular Biology over the past twenty years, an exponential growth has occurred in the number of laboratories and researchers that now find themselves working in the field of protein science. The tools of recombinant DNA methods have been largely responsible for this intellectual growth spurt, and many excellent texts have appeared on the subject of these technologies to keep up with the ever growing demands of the research community. Of particular value to the bench scientists are texts aimed at providing detailed laboratory protocols, rather than details of the theoretical basis for many of these methods. Thus, the series of laboratory manuals that have emerged from the Cold Spring Harbor Laboratories, for example, have been especially well received by the Molecular Biology community. Surprisingly, however, until very recently there have not been comparable texts devoted to the analysis of proteins.

In the text that follows, I have tried to fill this perceived need by the protein community. This field is so diverse, however, that any attempt at a comprehensive treatment of protein science would require a multivolume collection. I therefore have made no attempt to be comprehensive here. Rather, I have chosen those methods that I consider most relevant to the generalist in protein science. Some methods that are more specialized, like amino acid sequence analysis, have been given only a cursory treatment here. This is in recognition of the fact that most general protein laboratories will not participate directly in these studies, but are more likely to submit samples for sequence analysis to specialized laboratories. Other methods, such as analytical ultracentrifugation, that provide a

wealth of information of protein hydrodynamic properties have not been covered in the interest of brevity.

The methods covered here are those that can reasonably be performed in a standard protein laboratory, without specialized equipment or expertise. The techniques are presented in a "cookbook" fashion, emphasizing the how-to approach to science, and de-emphasizing, as much as possible, the theoretical basis for these methods. This bias is based on my personal belief that this type of treatment is most greatly needed in industrial, academic, and government laboratories at present.

Robert A. Copeland

Wilmington, Delaware
January 20, 1993

Acknowledgments

I am grateful to the many colleagues with whom I have collaborated over the years, and who have added to my knowledge of protein science. I would especially like to thank Professor Howard Tager of the University of Chicago for all his help, friendship, and advice. I would also like to express my appreciation to my students at the University of Chicago for their good humor and hard work during the writing of this manuscript. I am grateful to the Department of Chemistry and Biochemistry of the University of Delaware, which graciously provided me a place to work during the final preparation of this work. Finally, I would like to thank my family—Nancy, Lindsey, and Amanda—for putting up with me during the numerous crises that arose while this book was being written.

Contents

1

Introduction to Protein Structure

Proteins (from the Greek word *protos*, meaning first) occur ubiquitously throughout all organisms, where they participate in a wide variety of biological functions. Proteins provide the structural rigidity *and* flexibility associated with external body parts such as skin, hair, and the exoskeletons of arthropods. By changing their structure they provide the basis for biological motions, such as muscle contraction. A large class of proteins, the enzymes, serve as biological catalysts, enhancing greatly the rates of chemical reactions that are vital to life (Copeland, 1992).

How do proteins accomplish this great diversity of function? The answer to this lies in the structure and the structural variation accessible to proteins. In this chapter we will describe the structural components that make up proteins, the amino acids, and regular patterns of three dimensional arrangements of these groups that are commonly found in proteins. Ultimately it is the chemical reactivities of the individual amino acids and their arrangement in space that uniquely identify the active conformation of a protein and determine its functional activity.

THE AMINO ACIDS

Nature has selected twenty compounds to use as the building blocks for all proteins and peptides. These twenty molecules share the common structure shown in figure 1.1, and are collectively known as the natural amino acids. At neutral pH all of the amino acids are zwitterionic, having both a positively charged amino group and a negatively charged carboxylate group (hence the name amino acid). Intervening between these

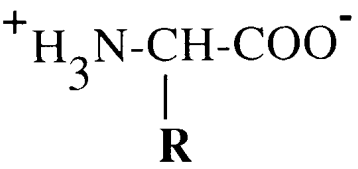


Figure 1.1 The general structure of an amino acid in its zwitterionic form.

charged functional groups is a single tetrahedral carbon atom, referred to as the alpha carbon. What distinguishes the amino acids from one another, both in terms of structure and chemical reactivity, is the identity of the substituent linked to the alpha carbon, which is denoted by R in figure 1.1 and is commonly referred to as the *amino acid side chain*. The side chain can be as simple as a hydrogen, in the case of glycine, or as complex as a fused bicyclic ring, in the case of tryptophan. Table 1.1 lists the twenty naturally occurring amino acids and illustrates the structures of their side chains. Also included in table 1.1 are the three-letter and one-letter abbreviations for the amino acids that are commonly used in the protein literature. The pK_a values for the amino, carboxyl, and side chain groups of each amino acid are also listed here. A quick perusal of this table will convince the reader that the pK_a values for the amino and carboxyl groups of all the amino acids are similar, hovering around 9.5 and 2.0, respectively. The pK_a values for the side chains, on the other hand, vary considerably (Dawson et al., 1969).

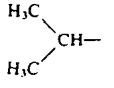
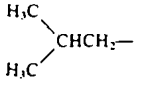
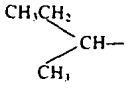
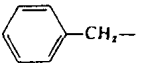
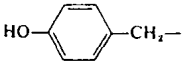
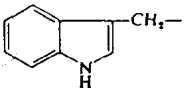
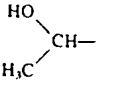
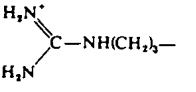
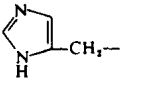
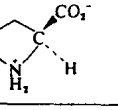
PEPTIDE FORMATION AND SEQUENCE NOMENCLATURE

Two or more amino acids can be covalently linked through a condensation reaction leading to an amide bond:

$$NH_3^+-CH(R_1)-COO^- + NH_3^+-CH(R_2)-COO^- \rightarrow$$
$$NH_3^+-CH(R_1)-C(O)-NH-CH(R_2)-COO^- + H_2O$$

The amide linkage formed in this process is called a *peptide bond*. When several amino acids are linked together in this fashion, the resulting polymer is called a *polypeptide*, and the individual amino acids along the polymer chain are called *residues*. One may ask what the difference is between a polypeptide and a protein. The difference is largely semantic. All proteins are, by definition, polypeptides. The term polypeptide is usually reserved for small polymers of, say, 50 or fewer amino acids.

Table 1.1 The structures of the naturally occurring amino acids found in proteins. This table is adapted from Dawson et al. (1969), with permission.

Amino acid (three- and one-letter codes, M_r)	Side chain R in $RCH(NH_3^+)CO_2^-$	pK_a's[a]
Glycine (Gly, G, 75)	H—	2.35, 9.78
Alanine (Ala, A, 89)	CH_3—	2.35, 9.87
Valine (Val, V, 117)		2.29, 9.74
Leucine (Leu, L, 131)		2.33, 9.74
Isoleucine (Ile, I, 131)		2.32, 9.76
Phenylalanine (Phe, F, 165)		2.16, 9.18
Tyrosine (Tyr, Y, 181)		2.20, 9.11, 10.13
Tryptophan (Trp, W, 204)		2.43, 9.44
Serine (Ser, S, 105)	$HOCH_2$—	2.19, 9.21
Threonine (Thr, T, 119)		2.09, 9.11
Cysteine (Cys, C, 121)	$HSCH_2$—	1.92, 8.35, 10.46
Methionine (Met, M, 149)	$CH_3SCH_2CH_2$—	2.13, 9.28
Asparagine (Asn, N, 132)	$H_2NC(=O)CH_2$—	2.1, 8.84
Glutamine (Gln, Q, 146)	$H_2NC(=O)CH_2CH_2$—	2.17, 9.13
Aspartic acid (Asp, *D*, 133)	$^-O_2CCH_2$—	1.99, 3.90, 9.90
Glutamic acid (Glu, E, 147)	$^-O_2CCH_2CH_2$—	2.10, 4.07, 9.47
Lysine (Lys, K, 146)	$H_3N^+(CH_2)_4$—	2.16, 9.18, 10.79
Arginine (Arg, R, 174)		1.82, 8.99, 12.48
Histidine (His, H, 155)		1.80, 6.04, 9.33
Proline (Pro, P, 115)		1.95, 10.64

A protein or polypeptide is uniquely defined by its amino acid composition, and by the order in which the amino acids occur along the linear chain of the polymer. This information is referred to as the *primary structure* or *amino acid sequence* of the protein. Note that no matter how many amino acids we string together to form a protein, we are always left with one terminal amino acid retaining its positively charged amino group, and the other terminal amino acid retaining its negatively charged carboxyl group. These residues are denoted the amino or N-terminus, and the carboxyl or C-terminus, respectively.

The individual amino acids within a polypeptide are indexed numerically in sequential order. The amino terminal residue is always designated number 1; one then continues numbering the residues that follow in ascending numerical order, ending with the carboxyl terminus.

When numbering the amino acids of a protein sequence, one always begins at the N-terminus and ends with the C-terminal amino acid.

This sequence nomenclature is illustrated in figure 1.2 for a peptapeptide.

PROPERTIES OF THE PEPTIDE BOND

When x-ray diffraction methods were first applied to small molecular weight amides, such as N-methyl acetamide, it was soon realized that the bond lengths observed in the crystal structures could not be accounted for by a typical carbonyl double bond (C=O) and a carbon-nitrogen single bond. Instead, both the carbonyl and carbon-nitrogen bond distances were intermediate between the known double and single bond distances for related compounds. To explain these results, delocalization of the π electron cloud over the triatomic O-C-N system is invoked (Creighton, 1984). This can be thought of in terms of the two resonance structures illustrated in figure 1.3.

Based on the crystallographically measured bond distances, the C-N

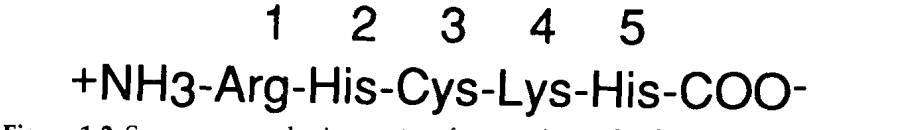


Figure 1.2 Sequence numbering system for proteins and polypeptides. Numbering begins with the N-terminal amino acid residue and proceeds in ascending sequential order ending with the C-terminal residue.

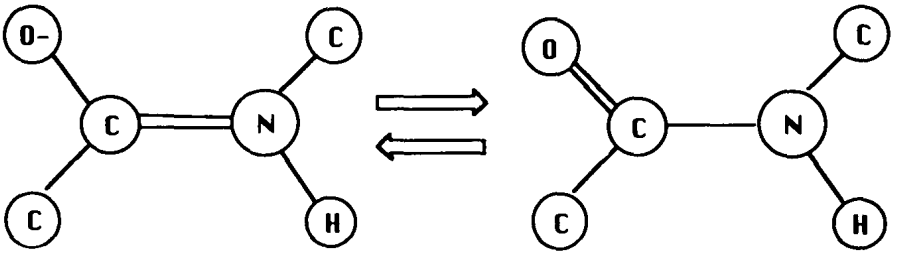


Figure 1.3 Resonance structures of the amide bonds of peptides.

bond appears to contain about 40% double bond character, and the C-O bond about 60% double bond character (Creighton, 1984). The significant π character of the C-N bond results in about a 20 kcal/mol resonance energy stabilization, and severe restriction of rotation about the C-N axis. Thus two peptide bond configurations are possible, cis and trans, with respect to the alpha carbons of the adjacent residues forming the peptide bond (see figure 1.4). In principle one could have polypeptides containing cis, trans, or both peptide bond isomers. However, the cis configuration leads to destructive non-bonding interactions that thermodynamically disfavor this conformation over the trans form by as much as three orders of magnitude. The vast majority of peptide bonds found in nature therefore occur in the trans configuration. One exception to this is prolyl peptide bonds. Occasionally, cis prolyl bonds have been observed in protein crystal structures, but even this is a rare event (Creighton, 1984).

Although rotation about the C-N bond is restricted, rotations about the C_α-N and C_α-C bonds can occur freely. The steric bulk of the amino acid side chain will, however, restrict to some degree the rotations about

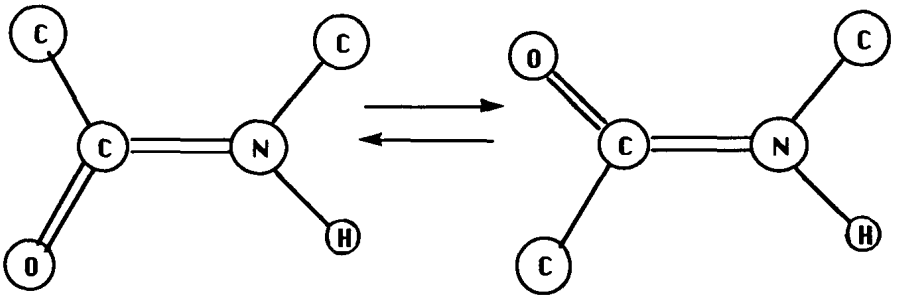


Figure 1.4 Cis and trans isomers of the amide bond. The trans form, shown on the left in this figure, is the thermodynamically favored form for all amide bonds found in proteins and polypeptides.

these bonds. One can define two dihedral angles to describe the orientation of peptide atoms with respect to these rotations, ψ and ϕ. These angles are illustrated in figure 1.5 for one peptide bond within a protein. In the late 1960's Ramachandran and coworkers surveyed the ψ and ϕ angles observed for amino acid residues within the crystal structures of proteins (Ramachandran and Sasisekhoran, 1968). Figure 1.6 illustrates the type of results one sees for such a survey for amino acids other than glycine. For glycine, the small size of the side chain (a proton) allows this amino acid greater freedom to survey extended ψ, ϕ space. What is most obvious from this plot is that the ψ and ϕ angles cluster around two sets of values. These two regions of high density correspond to the ψ and ϕ angles associated with two commonly occurring regular structural motifs that are found within proteins: the right-handed alpha helix, and the beta pleated sheet. These structural motifs are examples of protein *secondary structure,* an important aspect of the overall conformation of any protein (Branden and Tooze, 1991).

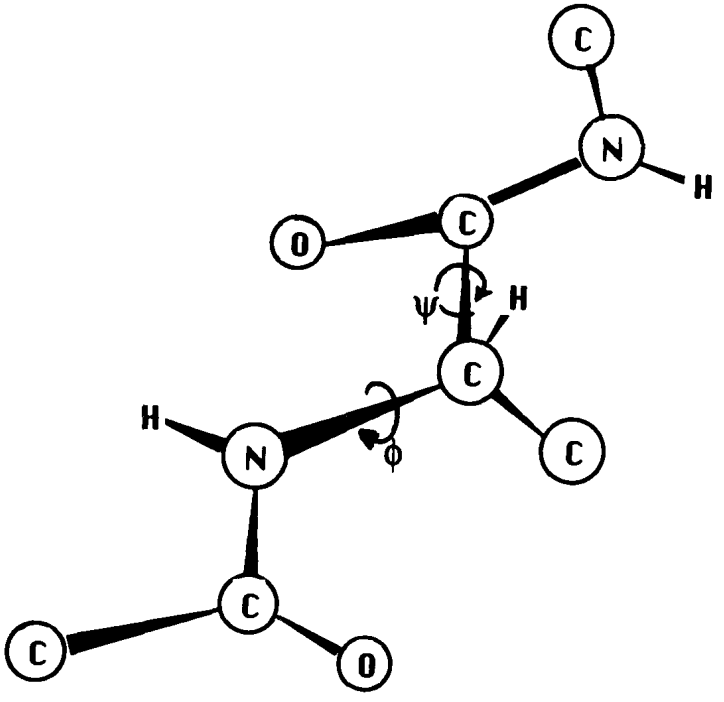


Figure 1.5 The dihedral angles of rotation for amino acids known as the Ramachandran angles, psi and phi.

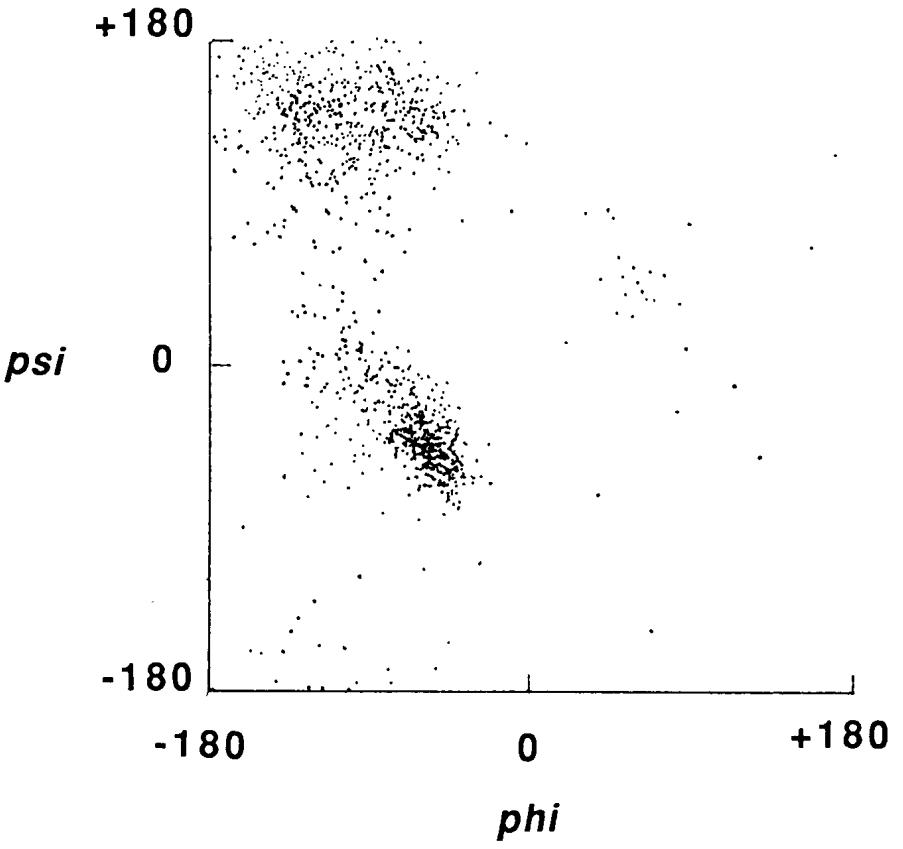


Figure 1.6 Ramachandran plot for the amino acid alanine illustrating the allowed combinations of psi and phi angles for this residue within proteins. Similar patterns are observed for all of the naturally occurring amino acids except for glycine. Because of the small size of the side chain of glycine (a proton), a wider range of allowed angle combinations is observed for this amino acid.

PROTEIN SECONDARY STRUCTURE

In the crystal structures of proteins one commonly finds regions of repeating structural patterns known as secondary structure. The two most commonly found secondary structures in globular proteins are the right-handed alpha helix, and the beta pleated sheet, which are illustrated in figure 1.7 (Pauling, 1960). These structural motifs are stabilized mainly by interamide hydrogen bonding. In the right-handed alpha helix, hydrogen bonding occurs between the backbone carbonyl of resi-

due i and the nitrogenous proton of residue i+3 along the polypeptide chain. This pattern of hydrogen bonding confers a very specific and regular structure to the alpha helix. Each turn of the helix requires 3.6 residues with a translation along the helical axis of 1.5 Å per residue, or 5.4 Å per turn.

The beta pleated sheet represents a fully extended configuration of the polypeptide backbone, in which hydrogen bonding occurs between residues on adjacent strands of the polypeptide. Note that the two strands forming the beta sheet could be two sections of the same contiguous polypeptide, or form the interface between two separate polypeptides. If we imagine a beta sheet within the plane of this page, both strands of the sheet could run in the same direction, for instance from C-terminus at the top of the page to N-terminus at the bottom, or they could run in opposite directions with respect to placement of the C- and N-termini. These structures are referred to as parallel and antiparallel beta sheets, respectively; both forms are found in proteins.

A third common secondary structure element in globular proteins is the beta turn. Beta turns are short segments (ca. five residues) of the amino acid sequence of the protein that allow the contiguous polypeptide to change direction. Because of the steric constraints imposed by such structures, residues making up turns most often contain small side chain groups. For this reason, glycine is commonly found associated with turns in globular proteins (Creighton, 1984).

Other regular secondary structures are sometimes found in proteins, such as the 3_{10} helix, poly-proline helices, and poly-glycine helices. These are, however, not common in globular proteins, and in many cases can be ignored (see Creighton (1984) for a discussion of these structures).

These regions of regular secondary structure are interspersed with sections of non-repeating unordered structure, commonly referred to as random coil. This does not imply that these regions of a protein are devoid of structure, but rather that the structure seen here is not repeating, or regular, nor does it fit into a specific category of secondary structure type. These regions of the protein are usually in greater dynamic flux than are the regions of regular secondary structure, and can thus play an important role in providing structural flexibility to the protein.

TERTIARY STRUCTURE

The next level in the hierarchy of protein structure is tertiary structure (figure 1.8). This term refers to how the elements of secondary structure

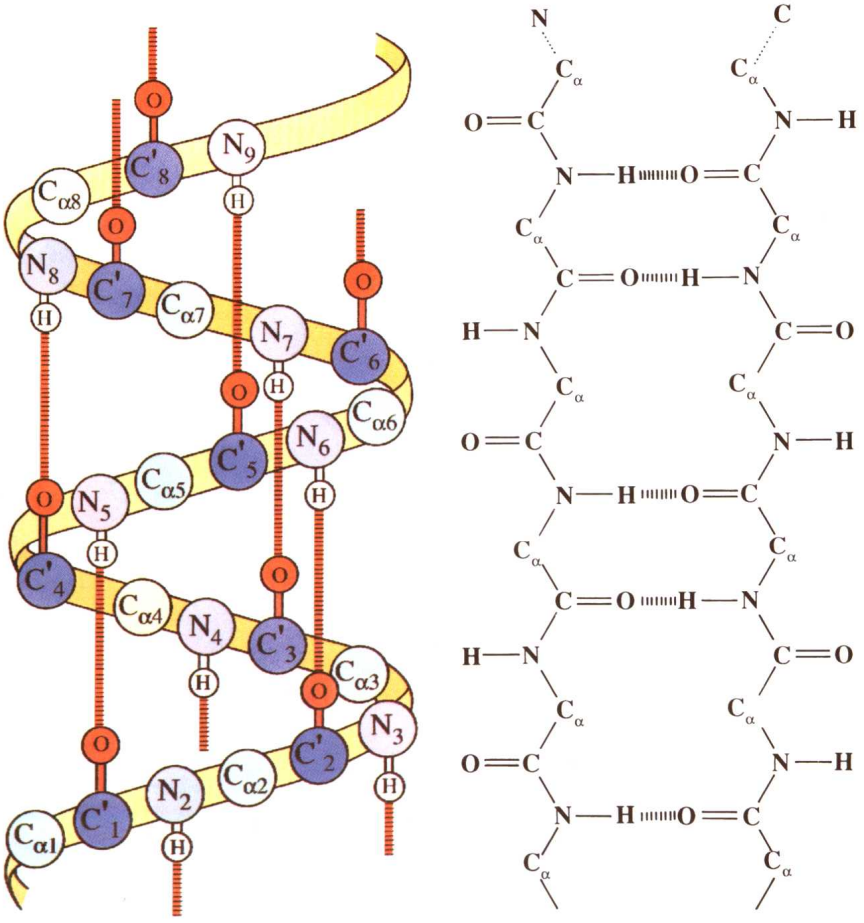


Figure 1.7 The right-handed alpha helix (left) and beta pleated sheet (right) secondary structures that are commonly found in globular proteins.

Figure 1.8 The hierarchy of protein structures: (1) primary structure or amino acid sequence, (2) secondary structure, (3) tertiary structure, and (4) quarternary structure.

arrange themselves in three dimensions in the folded protein, and how amino acid side chains interact with one another to form hydrogen bonds, electrostatic salt bridges, and hydrophobic-hydrophobic interactions (Branden and Tooze, 1991). Proteins fold into their native (i.e., naturally occurring biologically active) tertiary structure because of a variety of thermodynamic factors. Among these, one of the strongest driving forces is the need to shield nonpolar amino acids from the aqueous solvent into what has come to be known as the hydrophobic core of a protein. Likewise, folding favors interactions between polar amino acids and solvent molecules at the hydrophilic protein surface. Because of these types of forces, proteins often spontaneously fold into their native structures when presented with favorable solution conditions (Tanford, 1980).

Among the different side chain interactions that can occur in proteins is covalent bond formation through the oxidation of two sulfhydryl groups on cysteine residues. Such sulfur-sulfur bonds are known as disulfide bonds, and their presence in a folded protein provides an additional measure of stability to the native conformation. In multicysteine containing proteins, more than one arrangement of disulfide bonds may be possible. In such cases what almost always occurs is that only one specific set of disulfide bonds leads to an active protein. This is illustrated in figure 1.9 for the kringle 2 domain of human tissue plasminogen activator. Only the unique set of three disulfide bonds shown in figure 1.9 will lead to the correctly folded state of this protein (Vlahos et al., 1991).

The importance of protein tertiary structure cannot be overstated. It is the tertiary structure that gives a protein its overall shape and dimensions, and also provides a means of bringing into spacial proximity amino acid residues that may be distant in the linear sequence of the protein, but that need to come together to form the catalytic site of an enzyme, the binding pocket of a receptor, or a recognition site for the action of another protein. Thus tertiary structure is important for the following reasons:

- occlusion of hydrophobic residues from the polar solvent
- presentation of charged residues to solvent
- bringing groups that are distant along the sequence into close proximity for interactions
- providing the overall shape to the protein—important for establishing binding sites, sites for macromolecular recognition, etc