Protein Structure

Ivan Rayment
University of Wisconsin, Madison

I. Introduction
II. Amino Acids
III. Protein Structure Determination
IV. Structural Hierarchy
V. Secondary Structural Motifs
VI. Protein Stability
VII. Tertiary Structure
VIII. Membrane Proteins
IX. Superfamilies and Structural Evolution
X. Quaternary Structure
XI. Conclusions

GLOSSARY

Primary structure Linear sequence of amino acids in a polypeptide.
Quaternary structure Arrangement of polypeptides in macromolecular assembly.
Secondary structure A description of the three-dimensional structure adopted by a localized sections of the polypeptide chain.
Tertiary structure A description of the arrangement of secondary structural elements within the protein.
Units of length 1 Å = 0.1 nm.

PROTEINS are linear polymers of amino acids linked by amide bonds where the biological function is dictated by the sequence of amino acid residues within the polymer. Most proteins adopt a well-defined three-dimensional structure to fulfill their biological role. Thus, knowledge of protein structure is central to understanding the molecular basis of life.

I. INTRODUCTION

Proteins mediate the majority of biological processes. All proteins share the common feature that they are condensation polymers of amino acids whose sequence is specified by the genetic information contained within the genome of the organism. Complete DNA sequences for organisms ranging from Escherichia coli to humans suggest that the total number of proteins necessary for life lies in the range of 4200–50,000, although the number of genes in higher organisms is still under debate. Most of these proteins
adopt a well-defined three-dimensional structure in solution that is essential for protein function. Indeed unfolding or denaturation of a protein typically leads to a loss of biological activity.

The amino acid sequence of a protein contains all of the information necessary to dictate its final three-dimensional structure or fold. In many cases small proteins can be unfolded and refolded \textit{in vitro} without loss of activity. In more complex proteins, chaperones are frequently necessary to allow a protein to reach its properly folded or correct three-dimensional state. Chaperones in these instances recognize an incorrectly folded protein and provide an energetically favorable pathway, through the hydrolysis of ATP, for the protein to unfold and refold to reach its functional state. Even in these cases, the structure of the protein is dictated by its amino acid sequence.

In principle, it should be possible to deduce the structure of a protein from its amino acid sequence. At this time, it is not possible to perform \textit{ab initio} structure prediction with any great success. As such protein structure prediction remains one of the major problems in biology. Progress in structure prediction has been made through the combination of sequence and structural similarities. This offers hope that with the knowledge of sufficient structures across a wide range of organisms it should be possible to generate the structure of all unknown proteins. Although there is still much to be learned about protein structure, a series of fundamental features, folding rules, and structural motifs have been observed in many of the three-dimensional structures determined to date. These common features arise as a consequence of the amino acids used to build the proteins, the peptide bonds that join the amino acids, and the thermodynamic factors that control protein stability. These common threads in protein structure are described in the following.

II. AMINO ACIDS

All proteins are synthesized from the 20 $\alpha$-amino acids specified by the genetic code as shown in Fig. 1. The nature of an amino acid is determined by the “sidechain” attached to the $\alpha$-carbon (Table I). All of these amino acids, except for glycine which carries two hydrogens on its $\alpha$-carbon, have a chiral center located at the $\alpha$-carbon. Thus the amino acids exist as either the L- or D-isomers. Only the L-stereoisomer is utilized in protein biosynthesis (Fig. 2). This introduces chirality into all protein molecules that is the source of most of the asymmetric features found in protein structures. The use of only one of the two stereoisomers of the amino acids also establishes a structural uniqueness that is essential for biochemical specificity.

There are four classes of amino acids specified by the genetic code: (1) aliphatic amino acids, (2) aromatic amino acids, (3) polar amino acids, and (4) charged amino acids. These groups of amino acids provide the range of properties necessary to create a stable, functional folded protein.

As discussed elsewhere the primary driving force in protein folding and protein structure is the hydrophobic effect. This serves to sequestrate the hydrophobic side chains away from the bulk solvent. Once folded a typical protein is a densely packed entity that contains few holes larger than a water molecule. The aliphatic amino acids which include glycine, alanine, valine, leucine, isoleucine, and proline provide the range of small hydrophobic amino acids necessary to fill the gaps in the interior of the protein. Glycine and proline serve special roles in protein structure. Glycine is the smallest amino acid and is unique because it lacks a sidechain. This gives it more conformational freedom than any other amino acid. Glycine is often found in turns and loops where other amino acids would be sterically unacceptable. It is also found where secondary structural elements intersect and other side chains would introduce molecular collisions. In contrast proline is unusual because it is conformationally restricted. As such it is often found in turns since it introduces an inherent kink in the polypeptide chain without any entropic cost to protein folding. Proline is also unique in that it is the only amino acid (or technically an “imino acid”) that is commonly found to form a cis peptide bond between itself and the residue that precedes it in the polypeptide chain. In this instance the energy barrier to rotation is considerably less than all other peptide bonds (13 kcal/mol vs $\sim$20 kcal/mol). This post-translational conformational modification often represents a slow step in protein folding.

Phenylalanine, tyrosine, and tryptophan are large aromatic residues that are normally found buried in the interior of a protein and are important for protein stability. Tyrosine has special properties since its hydroxyl side chain may function as a powerful nucleophile in an enzyme active site (when ionized) and is a common site for phosphorylation in cell signaling cascades. Tryptophan has the largest side chain and is the least common amino acid in proteins. It has spectral properties that make it the best inherent probe for following protein folding and conformational changes associated with biochemical processes.

The polar amino acids include serine, threonine, cysteine, methionine, asparagine, and glutamine. These are an important class of amino acids since they provide many of the functional groups found in proteins. Serine often serves as a nucleophile in many enzyme active sites, and is best known for its role in the serine proteases. Both serine and threonine are sites of phosphorylation and glycosylation which are important for enzyme regulation and
cell signaling. Cysteine is the most reactive amino acid side chain. It serves as a potent nucleophile and metal ligand (particularly for iron and zinc), but is best known for its ability to form disulfide bonds, which often make an important contribution to the stability of extracellular proteins. Methionine is a fairly hydrophobic amino acid and typically found buried within the interior of a protein. It can form stacking interactions with the aromatic moieties of tryptophan, phenylalanine, and tyrosine. Asparagine and glutamine are close relatives of aspartate and glutamate but differ in the lack of charge and altered hydrogen bonding characteristics. In general these are not very reactive residues; however, asparagine is a common site for glycosylation.
TABLE I  Properties of the Amino Acids

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Side chain</th>
<th>Side chain pKₐ</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
<td>7.2</td>
</tr>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
<td>7.8</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
<td>6.6</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
<td>9.1</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
<td>5.3</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
<td>5.2</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
<td>3.9</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
<td>10.5</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Tyr</td>
<td>W</td>
<td>1.4</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
<td>13</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
<td>13</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
<td>8.4</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
<td>2.2</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
<td>4.3</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
<td>4.3</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
<td>10.5</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
<td>12.5</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
<td>6.0</td>
</tr>
<tr>
<td>Aspartate</td>
<td>Asp</td>
<td>D</td>
<td>3.9</td>
</tr>
<tr>
<td>Glutamate</td>
<td>Glu</td>
<td>E</td>
<td>4.1</td>
</tr>
</tbody>
</table>

The charged amino acids include aspartate, glutamate, lysine, arginine and histidine. As a group these amino acids are relatively abundant and are important for making proteins soluble. Thus, these residues are generally located on the surface of the protein unless they play a specific biological role. Aspartate and glutamate are negatively charged amino acids. Both of these residues can function as general acids or bases in enzyme catalyzed reactions. Likewise they are important metal ion ligands. Lysine, arginine, and histidine can carry a positive charge. Of these, arginine is constitutively positively charged since its pKₐ lies around 12.5. Lysine also plays an important role in coordinating negatively charged ligands; however, it functions as a nucleophile in some enzyme catalyzed reactions. Histidine is perhaps the most common and versatile catalytic residue in proteins. Its pKₐ of ~6.0 allows it to function both as a catalytic acid or base at physiological pH depending on its local environment. Histidine also has the ability to form covalent intermediates during catalysis such as phosphohistidine. In addition, it is often a ligand for transition metal ions such as iron and zinc.

A. Post-Translational Modifications

Once synthesized and folded, many proteins undergo post-translational modifications before they reach a functional state. Over 200 variant amino acid residues have been identified in proteins thus far. These changes are almost always achieved through an enzymatic pathway. The simplest changes include the formation of disulfide bonds (discussed later) and proteolytic processing of the polypeptide chain to yield a functional protein. Examples of proteolytic processing include the removal of signal peptides, the activation of zymogens to generate active forms of many proteolytic enzymes, and the maturation of viral proteins. Additionally proteolytic processing occurs in the biosynthetic pathway of many hormones. Other simple changes include the glycosylation of asparagine, serine, threonine, and phosphorylation of serine and tyrosine.

It is noteworthy that many post-translation modifications are associated with a sequence motif such that it is frequently possible to identify potential sites directly from the amino acid sequence. This arises because most post-translational modifications are the result of enzymatic pathways, which are usually highly specific. Thus protein sequences inferred from DNA sequence are often annotated with sites for post-translational modification. These sites should be viewed with caution since proof of

![Stereoisomers of L-alanine and D-alanine](image)

**FIGURE 2** Stereoisomers of L-alanine and D-alanine.
modification can only be obtained through chemical or physical experimentation.

B. Cofactors

Although many proteins derive all of their function from their constitutive amino acids, a large number of proteins require additional cofactors in order to fulfill their biological role. These cofactors provide chemical properties that are not present in the 20 amino acid residues. For example, none of the amino acids are capable of facilitating an oxidation/reduction reaction. A wide range of cofactors are utilized including inorganic ions such as Fe$^{2+}$, Mg$^{2+}$, and Zn$^{2+}$, or complex organic molecules that are normally described as coenzymes such as flavin adenine dinucleotide or nicotinamide adenine dinucleotide. If the coenzyme is covalently bound to the enzyme it is often called a prosthetic group. The complete enzyme is called a holoenzyme, whereas the protein in the absence of its cofactors is called an apoenzyme or apoprotein. Many apoproteins are considerably less stable in the absence of their cofactor. This suggests that although the amino acid sequence dictates the overall three-dimensional structure, the cofactor is an integral part of the protein.

C. Context Determines Function

Typically proteins and especially enzymes contain only a few residues that are absolutely vital for function. In contrast there are usually many other residues of the same type in the protein that do not fulfill any special role. In many cases the catalytic residues have different chemical and physical properties from the same amino acid in solution; for example, the $pK_a$ of the side chain might be several pH units higher or lower than the free amino acid. It is generally found that the behavior of an amino acid is profoundly influenced by the context of that amino acid within the protein. Altering the chemical properties of a functional group is one of the major attributes of protein structure and appears to be essential for the activity of most enzymes. There are many ways in which this is achieved; however, a simple example is placement of a charged residue in the interior of a protein such that the deionized state is favored. This serves to raise the $pK_a$ of aspartate and glutamate and lower the $pK_a$ of lysine.

III. PROTEIN STRUCTURE DETERMINATION

A major requirement for understanding protein structure is a large database of three-dimensional structures. This is particularly important for the comparative method of structure prediction. Although considerable progress has been made in recent years toward establishment of a comprehensive structural database many more protein models are needed before structures can be predicted with a high degree of confidence. There are two methods by which protein structures can be determined: X-ray crystallography and NMR. These techniques are complementary, with each having its advantages for providing information about specific aspects of protein structure. A detailed description of these methods is beyond the scope of this summary, but a few comments are noteworthy.

A. X-Ray Crystallography

The first structure of a protein, myoglobin, was determined by X-ray crystallography in 1958 and was followed soon thereafter by the structure of hemoglobin. At that time protein structure determination was a daunting undertaking and few structures were determined in the ensuing years. Fortunately continual developments in the fundamental understanding of X-ray crystallographic theory, data collection, and computational methods have made the determination of protein structure routine. The result of this approach is an electron density map, which is interpreted in terms of a molecular model. The strength of this technique is that it can be applied to any macromolecular assembly that can be crystallized. The overwhelming majority of structures in the protein database have been determined by X-ray crystallography.

The limiting factor in a successful X-ray structure determination is the growth of high quality crystals. In general if suitable crystals can be obtained a three-dimensional structure will be determined. The final quality of an X-ray structure is directly dependent on the three-dimensional order of the crystals since X-ray crystallography is an imaging technique. This is usually indicated by the “resolution” of the data. Resolution refers to the minimum diffraction spacing included in the structural determination where a smaller the number corresponds to a better structure. Typically a structure at 2.8 Å resolution is satisfactory to determine the path of the polypeptide chain, but data better than 2.5 Å are required to define the hydrogen bonding pattern in a protein with great confidence.

The one concern leveled at X-ray structures is the influence of the crystalline lattice on the observed conformation of the protein. Fortunately it has been demonstrated repeatedly that the structures of proteins observed in crystalline lattice are consistent with most of the biochemical measurements on the same protein. This arises because protein crystals typically contain about 50% solvent such that very little of a protein molecule is in contact with its neighbors in the crystal lattice and the packing forces are thermodynamically small. In some cases proteins are
enzymatically active in the lattice. In others conformational changes are observed between the substrate-free and substrate-bound forms of the enzyme. Typically this requires the crystallization of site-directed mutant proteins complexed with the substrate(s) or the study of complexes with substrate analogs. Except for the use of Laue techniques, protein crystallography yields a time-averaged view of the protein structure. Careful analysis of accurate X-ray diffraction data may provide some indication of conformational flexibility, but that aspect of protein structure is best suited to spectroscopic techniques such as NMR.

B. NMR

The use of NMR to determine protein structures is a more recent development than X-ray diffraction. It has the advantage that the analysis can be performed in the solution state of the protein which removes any artifacts introduced by crystallization. Its major disadvantage is the size limitation, which restricts most analyses to smaller proteins (<40 kDa), although it is anticipated that improvements in the technology will extend the size limitation.

Structural studies on proteins became possible with the advent of multidimensional NMR techniques. These rely on the use of isotopic labeling with $^{13}$C, and $^{15}$N and techniques to provide a facile method for assigning all of the $^1$H resonances in a protein, which would otherwise be a difficult task. The measurement of nuclear Overhauser effect (NOE) intensities provide much of the distance information necessary to derive a structure, although additional chemical shift information is needed for a high-resolution structural determination.

Once a set of distance information has been obtained a series of models are generated and optimized by energy minimization and molecular dynamics within the restraints imposed by the distance information. The advantage of this approach is that it provides structural information on the protein in solution, the drawback is that surface residues and loops appear less well defined because there are generally fewer distance restraints for these components. The great strength of NMR is that it can yield specific information concerning the $pK_a$ of an individual group in a protein as well as providing insight into the dynamical properties of the macromolecule.

IV. STRUCTURAL HIERARCHY

The structure of a protein is generally understood in terms of an organizational hierarchy that consists of protein sequence, local secondary structure, tertiary structure, and finally quaternary structure. The study of protein structure in these terms has led to a greater understanding of the underlying physical principles that control the conformation and function of proteins. This hierarchy also reflects one conceptual view of protein folding, where the local secondary structural elements form first, followed by the tertiary and quaternary structure. The following discussion of protein structure is organized according to the preceding hierarchy.

A. Primary Structure

The character of a protein is determined by the amino acid sequence and composition of the polypeptide chain. By convention the order of amino acids in a protein is listed starting at the N-terminal and ending at the C-terminal amino acid residue. The N-terminal amino acid carries a free amino group, whereas the C-terminal residue retains a free carboxyl group. These terminal residues of the polypeptide chain are also referred to as the amino and carboxy terminus of the protein, respectively. Almost all protein sequences are determined indirectly by DNA sequencing. Chemical sequencing, either by automated Edman degradation or by mass spectroscopy, is still necessary to identify a protein from its original source and to prove the presence of post-translational modifications. All sequences of interest should be examined for errors by resequencing and comparison with orthologous proteins.

B. Conformational Restrictions on the Polypeptide Chain

The amino acids in a protein are linked by an amide linkage that is referred to as the peptide bond (Fig. 3). There are several key features to this bond. It is planar as a consequence of the partial double bond character of the carbon–nitrogen bond. It is almost always in the trans configuration. The peptide bond is fairly rigid where the barrier to rotation is ~20 kcal/mol. The carbonyl oxygen and amide hydrogen carry a partial negative and positive charge, respectively, which allow each of them to form a hydrogen bond. This linkage profoundly influences the stability, conformation, structure, and function of proteins.

The atoms that form the backbone of the linear polypeptide chain are usually referred to as the main-chain atoms. By convention the conformational (torsional) angles adopted by the main chain atoms are denoted by $\phi$, $\psi$, and $\omega$ as shown in Fig. 3. Of these, $\omega$ describes the peptide bond and usually adopts a value of 180°. In principle there is free rotation about the other two angles ($\phi$ and $\psi$); however, the peptide bond and presence of a $\beta$-carbon places substantial restrictions on these conformational angles.

Conformational energy calculations and experimental observations based on high-resolution X-ray structure determination show that generally only ~8% of the possible
combinations of $\phi$ and $\psi$ are strictly allowed, whereas more generous energy considerations include a total of 20% (Fig. 4). These allowed regions of conformational space fall into three areas that are occupied by the major secondary structural motifs observed in protein structures. As indicated these belong to right- and left-handed $\alpha$-helices and $\beta$-sheet.

V. SECONDARY STRUCTURAL MOTIFS

The major three-dimensional motifs found in proteins were predicted to exist by Cory and Pauling in 1951 before the first protein structure determination through their study of the structures of small peptides. They recognized that secondary structural motifs must accommodate the hydrogen bonding potential of the peptide bond as well as utilize the conformational angles found in model peptides. This emphasizes the importance of hydrogen bonds in specifying the conformation of the polypeptide chain. In general every potential hydrogen bond donor and acceptor in a protein participates in one or more hydrogen bonds. This requirement explains the common occurrence of the $\alpha$-helix and $\beta$-sheet.

A. $\alpha$-Helix

The first secondary structural element predicted and identified in a protein was the $\alpha$-helix (Fig. 5). In the $\alpha$-helix the path of the polypeptide chain follows a right-handed arrangement where carbonyl oxygen of residue $i$ interacts with the amide hydrogen on residue $i + 4$. There are 3.6 residues per turn with a helical rise of 1.5 Å per residue which gives a helical pitch of 5.4 Å. As a consequence the side chains extend away from the helix axis every 100°. The side chains also extend toward the N-terminus of the $\alpha$-helix due to the chirality of the amino acids. This is a very compact arrangement of residues that satisfies the hydrogen bonding requirements of the polypeptide chain except for the four amide hydrogens at the N-terminus and four carbonyl oxygens at the C-terminus of the helix. The lack of hydrogen bonding at the ends of an $\alpha$-helix explains why they always contain more than one turn.

The length of $\alpha$-helices varies enormously from a few turns in globular proteins to hundreds as seen in extended proteins such as myosin. Keratin is one of the most abundant fibrous proteins and is almost entirely $\alpha$-helical in nature. In general there are some preferences for those amino acid residues found in $\alpha$-helices that prove useful for qualitative structure prediction. For example, proline is commonly found at the N-terminus of a helix, but rarely found in the middle. Not only does the proline lack an amide hydrogen but also the pyrrolidine ring restricts the preceding residue from adopting the conformational angles necessary for helix formation.

As indicated in the Ramachandran plot (Fig. 4), the left-handed $\alpha$-helix is an allowed conformation. It is observed
FIGURE 4 Ramachandran plot for a high-resolution protein structure (human UDP-galactose 4-epimerase). This reveals that only a limited part of conformational space is occupied by the main-chain torsional angles. The space enclosed by the solid lines represents the most energetically favorable regions of conformational space. The dashed lines indicate more generously allowed regions.

occasionally in proteins, but not for an extended number of residues on account of unfavorable interactions between side chains. Residues that adopt this conformation are usually located in turns. The dominance of the right-handed helical conformation over the left-handed is a direct consequence of the L-amino acids.

One other helical conformation is found in globular proteins: the $3_{10}$ helix. This differs from the $\alpha$-helix by the location of the hydrogen bond. In the $3_{10}$ helix a hydrogen bond is formed between the C=O$_{i}$ and H-N$_{i+3}$. The packing of the main-chain atoms in the $3_{10}$ helical conformation is quite tight thereby yielding nonlinear hydrogen bonds and thus is not found for extended periods. The only location that $3_{10}$ helical conformation is fairly common is at the C-terminal ends of $\alpha$-helices where it serves to terminate the helix with a tight turn. Three residues in the $3_{10}$ helical conformation constitute a type III turn. In principle the $\pi$-helix which has a hydrogen bond between C=O$_{i}$ and H-N$_{(i+5)}$ is an allowed conformation. It is not observed in proteins because the packing of the main-chain atoms would be too loose giving rise to a hole through the center of the helix. In addition there is steric hindrance between the side chains of adjacent residues along a $\pi$-helix.

B. Collagen Helix

Collagen is the most abundant protein in mammals. It is a fibrous protein that exhibits a helical repeat that is different from that of any of the previous conformations. This protein is characterized by a repeating motif $(\text{Gly-}X-\text{Y})_{n}$ where $X$ is usually proline and $Y$ is often 4-hydroxyproline. Each collagen molecule contains $\sim$1000 amino acid residues and is about $3000 \text{Å}$ long. This protein is synthesized as a preprotein that includes 200 additional residues at both the N- and C-termini that fold to form globular domains and serve to prevent the molecules
from assembling. After post-translational hydroxylation of approximately one half of the prolines and removal of the globular extensions at both termini each collagen pro- tomer adopts a left-handed polyproline (II) helical conformation and assembles with two other molecules to form a right-handed triple-helix (Fig. 6). Each strand of collagen is highly extended with 3.3 residues per turn and a translation of 2.9 Å per residue. Consequently individual strands are unstable and must aggregate for stability. The noninteger number of residues per turn is to accommodate the right-handed superhelix that contains 10 triplets per turn. With this arrangement, every third residue lies at the helix axis in such a way that only a glycine would fit (Fig. 6). This explains the (Gly-X-Y)ₙ sequence motif. Interestingly there is extensive hydrogen bonding in the collagen superhelix; however, it occurs between the carbonyl oxygen on one strand with a NH on a neighboring molecule rather than within the polypeptide as seen in the α-helix. The use of proline and hydroxyproline restricts the conformational angles available to the polymer and serves to stabilize the collagen fibers. Hydroxylation of the proline increases the stability of the fibers, although the exact reason for this enhanced stability is still unclear.

C. β-Sheet

The second most common and identifiable secondary structural conformation is the β-strand. In contrast to the α-helix, the polypeptide chain in a β-strand is almost completely extended with a translation per residue of 3.4 Å. An isolated strand is unstable because there are no interactions between residues that are close in sequence. Thus the
$\beta$-strand is only observed in conjunction with other strands where it can form complementary hydrogen bonds with opposing peptide groups. These strands can either associate in a parallel or antiparallel manner to form $\beta$-sheets. Association of multiple strands gives rise to larger sheets that may be built from a variety of antiparallel and parallel strands; however, there is a strong tendency to prefer structural motifs that are dominated by mostly parallel or antiparallel strands.

In both antiparallel and parallel strands the peptidyl oxygen and amide hydrogen form almost ideal hydrogen bonds with neighboring strands; however, the geometry is somewhat different in each type of sheet (Fig. 7). Both arrangements lead to stable structures; however, antiparallel $\beta$-sheets are generally considered to be more stable than sheets built solely from parallel strands. As discussed later parallel sheets typically are buttressed on both sides by additional layers of secondary structure where these are usually $\alpha$-helices. In contrast anti-parallel $\beta$-sheets often only require one additional layer of secondary structural elements to establish a stable fold.

The side chains in both parallel and antiparallel $\beta$-sheets extend alternatively to opposite sides of the sheet. Consequently the groups on adjacent residues within a strand do not contact each other. Rather there is considerable interaction between side chains on adjacent strands.
such that complementarity is observed. In addition, all β-sheets have a characteristic right-handed twist when viewed along the strand. This twist is considered to be the consequence of the interactions of the side chains with the backbone of the polypeptide chain and is thus a direct result of the chirality of the amino acids. The magnitude of the twist is somewhat variable but is usually more pronounced in antiparallel β-sheets.

There is a considerable contrast between the nature and usage of the α-helix and β-sheet. α-Helices are self-contained secondary structural elements that may contain a substantial number of amino acid residues even in globular proteins. By comparison, β-strands typically contain three to six amino acid residues and require an adjacent strand to form a stable folding unit. Proteins that are built from α-helices usually have a very high percentage of their
residues in the helical conformation (~80%) with comparatively few devoted to connecting regions. In proteins that are dominated by β-strands, typically <50% are in the β-conformation. This occurs because for every three to six residues in each strand there must be an equivalent number of amino acids devoted to a turn to bring the polypeptide chain back into a position where it can hydrogen bond to the same or neighboring β-strand. This emphasizes the importance of turns in protein structures.

D. Turns and Random Coil

Many proteins contain secondary structure that cannot be described as either helix or turn. This is typically classified as turn, loop, or random coil. These sections of the polypeptide chain are characterized by nonrepetitive conformational angles; however, this does not necessarily imply that these residues are less stable or less well ordered than the regular secondary structural elements. Many active site residues and components critical for ligand recognition reside in loops or random coil and adopt an exquisitely well-defined conformation.

On average, one third of all residues in proteins are involved in turns that serve to reverse the direction of the polypeptide chain. These turns are an essential feature of globular proteins and are almost always located at the surface. In contrast to α-helices and β-strands which have repetitive conformational angles, the conformational angles observed in turns occur in sets that are characteristic of each type (Table II). Turns have been classified according to the commonly observed groups of conformational angles and the number of residues involved. Of these the β-hairpin or reverse turn is the most common. This type of turn is frequently used to connect antiparallel β-strands.

Three general types of reverse turn have been described; types I, II, and III, which all contain four amino acid residues and normally exhibit a hydrogen bond between C=O(1) and H-N(i+3) (Fig. 8). Of these, the type III turn consists of a short section of residues in the 310 helical conformation. Additional variants of the type I and II class are observed in the I’ and II’ turns. These exhibit conformational angles for the central two residues of the turn that are the mirror image of types I and II. The observed conformation angles favor the presence of certain amino acids at specific locations in the turns. For example, glycine predominates at position (i + 3) and proline predominates at position (i + 1) in both types I and II turns. In all turns the central two amino acid residues do not form peptide hydrogen bonds within the turn itself and thus must either accommodate their hydrogen bonding potential via a side chain interaction with a neighboring residue or through interactions with the solvent. Thus polar or charged residues (Asp, Asn, Ser) are often located at the first residue of the turn so that they can form a hydrogen bond to the amide hydrogen of residue (i + 2). The need to satisfy the hydrogen bonding potential of the main chain atoms accounts for the placement of most turns at the surface of the protein.

VI. PROTEIN STABILITY

The term protein stability refers to the energy difference between the folded and unfolded state of the protein in solution. Remarkably, the free energy difference between these states is usually between 20 and 80 kJ/mol, which is of the magnitude of one to four hydrogen bonds. Although this suggests that proteins are only marginally stable, the stability is sufficient to prevent spontaneous unfolding at normal temperatures.

<table>
<thead>
<tr>
<th>TABLE II Conformational Angles of the Major Secondary Structural Elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>Secondary structure</td>
</tr>
<tr>
<td>Helical conformations</td>
</tr>
<tr>
<td>α-Helix</td>
</tr>
<tr>
<td>310 Helix</td>
</tr>
<tr>
<td>Collagen Helix</td>
</tr>
<tr>
<td>β-Strands</td>
</tr>
<tr>
<td>Antiparallel</td>
</tr>
<tr>
<td>Parallel</td>
</tr>
<tr>
<td>β-Turns</td>
</tr>
<tr>
<td>Type I</td>
</tr>
<tr>
<td>Type I’</td>
</tr>
<tr>
<td>Type II</td>
</tr>
<tr>
<td>Type II’</td>
</tr>
<tr>
<td>Type III</td>
</tr>
</tbody>
</table>
Protein stability is determined by an enormous number of weak interactions observed in the folded state which have to be balanced against an almost equivalent set of interactions with water in the unfolded state. This is a complex problem since every amino acid residue has potential for polar interactions via the peptide bond and a variety of ionic, polar, and nonpolar interactions through its side chains. This accounts for the difficulty in predicting structure directly from its amino acid sequence since the errors in any energy computation are far larger than the net stability of the protein.

A. Hydrophobic Effect

The major driving force in protein folding is the hydrophobic effect. This is the tendency for hydrophobic molecules to isolate themselves from contact with water. As a consequence during protein folding the hydrophobic side chains become buried in the interior of the protein. The exact physical explanation of the behavior of hydrophobic molecules in water is complex and can best be described in terms of their thermodynamic properties. Much of what is known about the hydrophobic effect has been derived from studying the transfer of hydrocarbons from the liquid phase into water; indeed the thermodynamics of protein folding closely follow the behavior of simple hydrophobic molecules in water.

Studies of hydrocarbon models demonstrate that at room temperature the insolubility of hydrophobic compounds is dominated by entropic rather than enthalpic considerations. This is often explained as a water-ordering effect where the insertion of a hydrophobic molecule into an aqueous environment induces a diffuse, liquid-like water shell surrounding the molecule, akin to the formation of clathrates around noble gases and simple hydrocarbons. This shell forms because the hydrophobic compound cannot form hydrogen bonds to the water that surrounds it. Consequently those water molecules have a more restricted set of neighbors with whom to fulfill their hydrogen bonding capacities. This reduces their degrees of rotational freedom and thus leads to a reduction in entropy. This simple explanation based on ordering of water is
inadequate to fully explain the behavior of hydrophobic molecules in solution since both the standard changes in the entropy and enthalpy for the transfer of a hydrophobic molecule to water are strongly temperature dependent. Surprisingly the value for the free energy change for the transfer of a hydrocarbon to water is rather temperature insensitive as a consequence of compensatory changes in the entropy and enthalpy. These temperature dependencies are a consequence of the large temperature-insensitive heat capacity of hydrophobic molecules in solution. The water-ordering effect appears to be the source of the anomalously high heat capacity of hydrophobic compounds in water.

Although there is no general agreement on the molecular basis of the hydrophobic effect, there is a good correlation between free energy of transfer of an organic molecule to water and its hydrophobic surface area, whereas there are no general correlations between the molecular features of the solute and entropic and enthalpic changes. Many studies have shown that the transfer of 1 Å² of hydrophobic area to water is accompanied by an unfavorable increase in the free energy of 80–100 J/mol. The fact that this correlation is found to be fairly independent of the nature of the hydrophobic solute clearly indicates that the hydrophobic effect is a fundamental property of water.

Studies on the heat capacities changes observed at the protein folding transition show that protein denaturation is analogous to the transfer of hydrophobic molecules to water. Furthermore it is well established that the stability of a protein is directly proportional to the difference between the exposed hydrophobic surface area in the unfolded and folded state. In recent years, site-directed mutagenesis has demonstrated the same thermodynamic relationship between the hydrophobic buried surface area and stability as observed for the transfer of organic hydrocarbons into water. That is, each buried 1 Å² of hydrophobic surface area contributes ~80 J/mol to the stability of the protein when the only difference is the change in surface area. Clearly the behavior of a protein in solution is more complex than that of a simple hydrocarbon. As noted earlier, proteins are only marginally stable. It would appear that the change in exposed hydrophobic surface area that accompanies protein folding slightly more than compensates for the decrease in entropy of the polypeptide chain as it adopts a well-defined conformation. This explains why it is so difficult to predict protein structure.

B. Hydrogen Bonds

Hydrogen bonds (D—H···A) are primarily electrostatic in nature and involve an interaction between a hydrogen attached to an electronegative atom (D—H) and another electronegative acceptor atom (A) that carries a lone pair of electrons. In biological systems the electronegative atoms in both cases are usually nitrogen or oxygen. The distance between the donor and acceptor atoms is usually in the range 2.8–3.1 Å where the D—H bond tends to be collinear with the lone pair of electrons. There is some variability in the geometry of the hydrogen bond, which is consistent with the predominantly electrostatic nature of this interaction. For example, in the α-helix and antiparallel β-sheet the N—H is approximately colinear with the C=O bond rather than be aligned with the lone pairs of the carbonyl oxygen. Many of the hydrogen bonds in proteins occur in networks where each donor participates in multiple interactions with acceptors and each acceptor interacts with multiple donors. This is consistent with the ionic nature of hydrogen bonds in proteins.

Originally it was believed that hydrogen bonds made an important contribution to protein stability on account of the extensive hydrogen bonding observed in α-helices and β-sheets. Indeed, virtually every hydrogen donor and acceptor in a protein are observed to form an interaction within the folded structure or to the external solvent. However, protein stability is the difference in free energy between the unfolded state and the folded state. In the unfolded state the polar components are able to form perfectly satisfactory hydrogen bonds to water that are equivalent to those found in the tertiary structure of the protein. Thus hydrogen bonding is energetically neutral with respect to protein stability, with the caveat that any absences of hydrogen bonding in a folded protein are thermodynamically highly unfavorable.

Although hydrogen bonds do not contribute to stability they are a major determinant of protein conformation. The necessity to form hydrogen bonds accounts for the α-helices and β-strands that abound in protein structures.

C. Disulfide Bonds

Many extracellular proteins contain disulfide bonds. In these proteins the presence of disulfide bonds adds considerable stability to the folded state where in many cases reduction of the cystine linkages is sufficient to induce unfolding. The source of the stability appears to be entropic rather than enthalpic. The introduction of a disulfide bond reduces the entropy of the unfolded state by reducing the degrees of freedom available to the disordered polypeptide chain. This stabilizes the folded state by decreasing the entropy difference between the folded and unfolded state. This suggests an obvious strategy for increasing the stability of a protein through the introduction of disulfide bonds. Although this might seem a simple task, the geometry of the disulfide bond is rather restricted. As a consequence the number of locations that can accommodate the replacement of two residues by cysteines in a protein
without reducing the thermal stability of the folded protein are quite limited. Studies on T4 lysozyme have shown that if suitable locations can be found, the degree of stability introduced into a protein is proportional to the size of the closed loop generated by forming a disulfide bond.

**D. Ionic Interactions**

The association of two oppositely charged ionic groups in a protein is known as a salt bridge or ion pair and is a common feature of most proteins. Typically these interactions contribute very little to protein stability since the isolated ionic groups are so effectively solvated by water. As a consequence very few unsolvated salt bridges are found in the interior of proteins. Furthermore, salt bridges are rarely conserved in orthologous proteins.

**E. Dipole–Dipole Interactions**

Dipole–dipole interactions are weak interactions that arise from the close association of permanent or induced dipoles. Collectively these forces are known as Van der Waals interactions. Proteins contain a large number of these interactions, which vary considerably in strength.

The strongest interactions are observed between permanent dipoles and are an important feature of the peptide bond. In the peptide bond the dipoles associated with the peptide carbonyl and amide group are aligned and give rise to a significant dipole moment (3.5 Debye units for a peptide bond versus 1.85 for a water molecule). These interactions fall off with the inverse of the second to third power when the dipoles are fixed and to the sixth power when they are free to rotate. So, for example, there is a substantial positive dipole at the amino-terminal end of an α-helix where the dipoles are constrained and aligned. As a consequence the N-terminal end of an α-helix is often utilized to bind negatively charged ligands in enzyme active sites.

Permanent dipoles may also induce a dipole moment in a neighboring atom or group. This is a stabilizing interaction, but is much weaker than that observed between permanent dipoles. These type of interactions are important since they change the charge distribution of neighboring atoms which in turn can profoundly influence activation barriers in enzyme catalyzed reactions.

London or dispersion forces are the weakest of all the dipole–dipole. These are best described in quantum mechanical terms, but may be viewed qualitatively as the consequence of the transient asymmetry in the charge distribution in a neutral atom that induces a favorable dipole in a neighboring neutral atom thus leading to a weak attraction. These forces are inversely proportional to the sixth power of the separation. Although these forces are very weak there are an enormous number present within a folded protein such that they contribute significantly stability of the folded state. As a group, the Van der Waals forces are important for stabilizing interactions between proteins and their complementary ligands whether the ligands are proteins or small molecules.

**VII. TERTIARY STRUCTURE**

**A. Protein Folding Rules**

Examination of a large number of protein structures has yielded a few common rules about the folds that proteins can adopt as listed in the following. The theoretical basis for these features is not well understood, but most appear to result from the chirality of the amino acids, entropic considerations, and the necessity to establish a hydrophobic core.

1. Secondary structural elements that are close in the sequence of a protein are often adjacent in the folded protein. It is less common to find secondary structural elements that are far apart in the sequence and close together in the structure. The exception to this arises where an auxiliary domain has clearly been inserted into a loop in a protein. This is probably the most entropically favorable way to arrange secondary structural elements within a folded protein.

2. Adjacent parallel β-strands are almost exclusively connected by right-handed crossovers (Fig. 9). It is believed that this feature arises from the chirality of the amino acids that leads to a net right-handed twist in the polypeptide chain.

3. There are no topological knots in proteins.

4. Proteins always contain more than one layer of secondary structural elements. This rule arises because proteins always contain a hydrophobic core formed by the association of hydrophobic side chains.

5. α-Helices and β-sheets typically associate in discrete layers of the same type of secondary structural elements. This feature is the consequence of the necessity to fulfill the hydrogen bonding requirements of the polypeptide chain and packing considerations. Typically the interior of a protein does not contain any holes larger than a water molecule. Because α-helices and β-strands differ greatly in their cross-sectional diameters, inclusion of these in the same layer would result in a poorly packed protein interior. In addition a mixture of α-helices and β-strands in the same layer would not fulfill the hydrogen bonding potential of the β-strand.
B. Folding Motifs

Many protein structures are dominated by a few simple folding motifs. These represent thermodynamically favorable arrangements of secondary structural elements. These include the $\beta\beta$, $\beta\alpha\beta$, and $\alpha\alpha$ motifs as illustrated in Fig. 10.

The $\beta\beta$ and $\beta\alpha\beta$ motifs are commonly used to connect antiparallel and parallel $\beta$-strands, respectively. The $\beta\beta$ motif is frequently connected by a hairpin turn, which provides a compact way of changing the direction of the polypeptide chain. In the same way, the $\beta\alpha\beta$ motif provides a compact module where the width of the $\alpha$-helix is similar to that of the combined width of the two $\beta$-strands. It also provides a hydrophobic core. The dimensions of the $\beta\alpha\beta$ motif explain why large parallel sheets that are built with this motif always have $\alpha$-helices on both sides since there is insufficient space on one side of a sheet to accommodate all of the connecting helices.

A variety of $\alpha\alpha$ motifs are found in proteins depending on whether the $\alpha$-helices are in contact with each other after the connecting loop. In cases where the $\alpha$-helices are in contact they are typically inclined at an angle of either 20 or 50° reflecting the optimal ways to interdigitate side chains at their intersection. Both types of interaction are abundant in proteins and give rise to parallel or crossed helical bundles. There are also many important examples of $\alpha\alpha$ motifs where the connections between the two helices are longer to create a ligand binding site. Important examples of this type of motif are the helix-turn-helix motifs found in calcium binding proteins and DNA binding proteins.

C. Domains

The tertiary structure of a protein describes the manner in which the secondary structural elements are arranged in three dimensions to create a stable molecular entity. In many cases it is convenient to describe a protein in terms of regions of the polypeptide chain that might fold autonomously. These regions are called domains and much of the discussion of tertiary structure centers on classification of these units of protein structure.

Domains in proteins take on many forms. On some occasions it is clear that domains are connected by flexible hinge regions and that the domains could be expressed independently. In other cases the domains are built from apparently distant segments of the protein sequence such that it would be difficult to express those domains without rearrangement of the DNA. This illustrates an important difference in the use of "domain" in structural and molecular biology, since in the latter the term usually indicates a linear section of DNA that appears to influence a biological property whereas in structural biology it represents an three-dimensional entity.

D. Protein Folds

Structural studies on proteins have uncovered a very wide variety of protein folds. At this time the upper limit of the number of unique ways in which proteins can fold is unknown; however, genomic sequencing has provided a limit for the maximum number of folds that might be needed for the life of an organism by providing an upper limit to the number of proteins in the genome. Fortunately, the number of unique folds is likely to be considerably less than the total number of proteins since many proteins of dissimilar function have been found to contain the same fold.

The assortment of protein folds observed thus far, at first glance, appears bewilderingly complex. Careful analysis of the common structural and topological features of these structures has lead to a classification of protein folds according to the content and arrangement of
the α-helices and β-strands. In turn, this has provided insight into the common underlying principles of protein structure. Several important databases exist of protein structures and tertiary structure classification. These include the RSCB (//www.rcsb.org/pdb/), CATH classification (//www.biochem.ucl.ac.uk/bsm/cath/), and SCOP (//scop.mrc-lmb.cam.ac.uk/scop/) structural databases. The first includes all of the coordinates for structures that have been made publicly available. The second two databases contain structural classifications for all the protein deposited in the RSCB. Both of these systems initially classify proteins into five major groups: all α, all β, α/β (where these secondary structural elements alternate through the fold), α + β (where the sections containing these secondary structural elements are segregated), and small proteins that are stabilized by metal ligands or disulfide bonds. Additional classifications have been added to incorporate multidomain proteins, membrane proteins, and peptides. Representative members of these families are discussed in the following.

E. All α-Proteins

Proteins that fall into this class typically consist of predominately α-helices (>60%), but may contain a small amount of β-sheet at their periphery (See CATH Classification). Historically the first two protein structures determined, myoglobin and hemoglobin, belonged to the all-α class.
These are representative members of the crossed α-helical bundle motif and demonstrate one of the effective ways of packing helices into a protein core (Fig. 11a). The four-helix bundle illustrated in Fig. 11b shows the second way in which helices associate. When the connecting loops are short the packing leads to an antiparallel arrangement of helices; however, mixtures of parallel and antiparallel helices are also observed in folds that contain longer intervening sequences.

F. All β-Proteins

There are a large number of diverse protein folds that fall into the all-β class, but they all contain ~50% β-strand with only a very small amount of α-helix. Several representative examples of these are shown in Fig. 12. In structures composed primarily of β-strands there is always more than one layer which is necessary to establish a hydrophobic core. This means that at the edge of the sheets there must be some form of compensation for the disruption of the hydrogen bonding pattern. Most of the folds in this class are formed from antiparallel arrangements of β-strands.

A large number of the all-β structures arrange their strands to form barrel-like or sandwich structure. The simplest of these arrangements is the up-and-down barrel or “clam motif” found in the retino binding superfamily of proteins (Fig. 12a) where each strand adds to the next in an antiparallel manner until the barrel is complete. In this group of proteins the interior of the barrel provides a binding site for hydrophobic ligands.

A substantial number of all-β proteins are built from antiparallel strands in which adjacent strands are not directly connected. Many of these contain a topological feature known as a Greek Key in which the first strand connects across the top of the barrel or sandwich to the fourth
FIGURE 12  Ribbon representations of typical all-ß proteins. (a) Retinol binding protein, (b) immunoglobulin fold as seen in the C8 Surface glycoprotein N-terminal domain, (c) pectate lyase, and (d) viral coat protein found in satellite tobacco necrosis virus.
strand which then returns via two hairpin connections to the strand adjacent to the first. This feature contains a handedness that is only observed in one sense as shown in Fig. 13a. Connection of two of these features gives rise to an eight stranded β-barrel; however, Greek Key motifs are utilized in many ways to form closed structures. An alternative way of forming a closed barrel is found in proteins that exhibit a “jelly roll” topology (Fig. 13b). This is an abundant motif that is commonly found in virus capsid proteins.

The inclusion of a few sections of random coil or the occasional α-helix into an all-β protein allows for the generation of some remarkable motifs as shown in the β-propellers and β-helical folds. These folds illustrate the versatility of the β-strand when the hydrogen bonding potential of the polypeptide chain is fulfilled.

G. α/β Proteins

The α/β class of proteins contains many of the folds that incorporate parallel β-sheets. These folds exist in two major subclasses: the first contains a closed circular β-sheet surrounded by α-helices which forms a barrel; the second is based on an open sheet typically surrounded on both sides by α-helices (Fig. 14). Both of these arrangements are abundant in biosynthetic enzymes.

1. TIM Barrel

Triosephosphate isomerase was the first enzyme shown to contain an (α/β)8 barrel and thus established this motif as the TIM barrel. This fold consists of eight parallel β-strands connected by right-handed helical crossovers and is one of the most common folds found in enzymes. The TIM barrel typically contains approximately 200 amino acid residues. Contrary to the appearance of the ribbon drawing (Fig. 14a), the interior of the barrel is closely packed by the side chains protruding from the β-strands. The strands are inclined at an angle of approximately 30° to barrel axis, which is necessary to allow efficient packing of the interior. The necessity to form a closely packed interior explains why these barrels are almost always formed from eight strands. There are several variations on the TIM barrel that include the addition and subtraction of β-strands as well as the introduction of antiparallel β-strands as observed in enolase. These variations attest to the versatility of this fold.

The active sites of triosephosphate isomerase and all other enzymes that contain to this fold are located at the C-terminal end of the β-strands. Typically the catalytic residues reside at the end of the strands and are distributed around the barrel. The loops that connect the strands to the α-helices normally provide the components necessary for substrate specificity. The length of these connecting
FIGURE 14  Ribbon representations of $\alpha/\beta$ proteins. (a) Triosephosphate isomerase, (b) dinucleotide binding domain of lactate dehydrogenase (c) alkaline phosphatase which is an example of a complex member of the $\alpha/\beta$ class of folds.
loops is enormously variable, whereas the length of the 
\( \beta \)-strands are similar in all enzymes.

2. Open \( \beta \)-Sheets

The second class of proteins in the \( \alpha/\beta \) family of folds contains a large open sheet formed from mostly parallel 
\( \beta \)-strands with helices on both sides. In contrast to the TIM barrel there are fewer limitations on the number of 
strands within the sheet and may vary from 4 to 10. The first example of this type of fold was seen in lactate dehydro-
genase which contains a motif that is widely observed in dinucleotide binding proteins (this motif is often referred to as the Rossmann fold) and was the first example of a domain superfamily (Fig. 14b). The observation of 
a common fold in the dehydrogenases by Rossmann and coworkers started the entire field of structural comparison 
and study of structural evolution.

All of the connections between \( \beta \)-strands are formed by 
right-handed crossovers. As a consequence, the strand 
order within the sheet must reverse in order to place helices 
on both sides of the sheet (Note: the consecutive strand 
order in the \( (\alpha/\beta)_S \) barrel places the \( \alpha \)-helices on one 
side of the sheet). In the classical Rossmann fold, which 
contains six \( \beta \)-strands, the N-terminal strand in the fold is 
located adjacent to the center of the motif. The first two 
\( \alpha \)-helices lie on one side of the sheet as the first three 
strands are added. Thereafter the chain returns to the cen-
ter of the sheet and adds the next three strands with the 
reverse strand order such that the subsequent helices are 
added on the opposite side of the sheet.

There are many varieties of open sheet \( \alpha/\beta \) proteins 
which include differing numbers of strands, connections 
between strands that are not adjacent and incorporation of 
antiparallel strands. In most cases the ligand binding sites 
are located at the C-terminal ends of the \( \beta \)-strands and lie at 
the crevice at the edge of the sheet where the strand order is 
reversed. The loops that connect the strands to the helices 
typically provide the residues necessary for specificity. 
The size of the connecting loops are enormously variable 
in \( \alpha/\beta \) proteins.

H. \( \alpha + \beta \) Proteins

The \( \alpha + \beta \) class of proteins is highly variable, indeed 
over a hundred distinct folds have been observed in this 
group. Members of this class typically contain one or more 
\( \beta \)-sheets which have a bias toward antiparallel 
connections. As such the \( \alpha \)-helical and \( \beta \)-sheet regions tend to 
be segregated along their sequences. Several examples of 
proteins that fall in this class are shown in Fig. 15. In the 
simplest cases the helices lie on one side of the sheet which 
may be comparatively flat or steeply curved as in ubiquitin.

These are known as \( \alpha/\beta \) folds. In more complex folds multiple 
layers of sheet and additional layers of helices have 
been observed to give rise to \( \alpha/\beta \) (as in ribonuclease) and 
\( \alpha/\beta/\alpha \) (as in glutamine phosphoribosyl pyrophosphate 
aminotransferase N-terminal domain).

I. Small Proteins, Unusual Folds

There are a substantial number of small proteins that defy 
classification into one of the groups listed above. Some of 
these have limited regular secondary structure whereas 
others are stabilized by metal ligands, cofactors, or disul-
hide bonds. Examples of these folds include the zinc–finger 
DNA binding motifs, many small iron–sulfur proteins, 
toxins and protein-inhibitors (Fig. 16).

VIII. MEMBRANE PROTEINS

Approximately one third of all proteins are tightly asso-
ciated with membranes. These are much more difficult 
to crystallize or study by NMR than water-soluble pro-
teins. As a consequence, there are far fewer structures of 
membrane proteins. Even so, those that have been deter-
mined provide insight into the manner in which polypep-
tide chains interact with lipid bilayers.

Membrane proteins fall into two classes: peripheral and 
integral. Peripheral membrane proteins are associated with 
the membrane, but may be removed by high concentra-
tions of salt or metal chelators such as EDTA. In most as-
pects the structures of peripheral membrane proteins are 
very similar to water-soluble proteins. Integral membrane 
proteins differ in that they are very difficult to extract from 
the lipid bilayer and require detergents for solubilization. 
Detergents disrupt the lipid bilayer and bind to the 
hydrophobic surfaces of the protein that are buried within 
the membrane.

Integral membrane proteins all share the common prob-
lem of inserting a polypeptide chain into the hydrophobic 
interior of the lipid bilayer. This poses a thermodynamic 
problem on account of the hydrogen bonding propensity 
of the polypeptide chain. Clearly any segment of the pro-
tein that passes through the lipid bilayer must accommo-
date the hydrogen bonding potential of the polypeptide 
chain. Originally it was believed that an \( \alpha \)-helix would be 
the only secondary structural element to pass through the 
lipid bilayer since it alone fulfills the hydrogen bonding 
capacity of the polypeptide chain in a consecutive man-
ner. Indeed the \( \alpha \)-helix is the only way to pass a single 
transmembrane segment of protein through a membrane. 
However, although a large number of membrane proteins 
are formed from \( \alpha \)-helical bundles a significant number are 
built from \( \beta \)-strands. Both of these strategies for building
integral membrane proteins present interesting biophysical problems.

A. α-Helical Membrane Proteins

Many fully inserted membrane proteins utilize bundles of α-helices to span the lipid bilayer. The structure of the photoreaction center, the first membrane protein whose structure was determined, clearly shows this strategy (Fig. 17a). It consists of polypeptide chains that span the lipid bilayer utilizing 11 transmembrane α-helices. Interestingly the surfaces that face the interior of the bilayer are more hydrophobic than the interior of the protein, whereas the components that face the aqueous environments are similar to the surfaces of water soluble proteins. Thus there is no tendency for proteins to unfold in the lipid bilayer. This suggests that the same forces that stabilize water-soluble proteins are responsible for the stability of membrane proteins. α-Helices are the major component of proton
pumps such as bacteriorhodopsin (Fig. 17c) and K⁺ channel proteins.

**B. β-Sheet Membrane Proteins**

The outer membranes of gram-negative bacteria contain channels that allow the diffusion of small solutes and ions into the periplasmic space. These channels are formed by bacterial porins that are built almost entirely of antiparallel transmembrane β-strands (Fig. 17b). The topology of these proteins is exceedingly simple, consisting of up-and-down strands where the first strand hydrogen bonds with the last strand of the sheet. In this way the hydrogen bonding potential of the β-strands is completely satisfied and a hydrophobic surface is presented by the side chains that extend into the lipid bilayer. The channel lies down the middle of the barrel. The side chains that line the pore provide some selectivity for the nature of the solutes that diffuse through the channel. All bacterial porins exist as oligomers (mostly trimers) where the interface between the porin subunits form a hydrophobic interior that is otherwise missing from these proteins.

This type of antiparallel packing of β-strands in integral membrane proteins has also been observed in hemolysin. Hemolysin is a heptameric pore forming protein from *Staphylococcus aureus*, where each of the subunits contributes two antiparallel β-strands to the transmembrane segment creating a 14 stranded barrel.

**C. Other Membrane Motifs**

Not all membrane proteins extend completely through the lipid bilayer, indeed many of the biosynthetic enzymes that are tightly bound to the membrane only extend into
FIGURE 17 Ribbon representation of selected integral membrane proteins. (a) Photoreaction center, (b) bacterial porin, (c) bacteriorhodopsin, (d) prostaglandin synthase.
one face. A good example of this is prostaglandin synthase (Fig. 17d). This enzyme converts arachidonic acid into prostaglandin (PGH₂) and exhibits two catalytic activities: a cyclooxygenase and peroxidase. It is an important enzyme since it catalyzes a critical step in the biosynthesis of a wide range of eicosanoids that control the inflammatory response, pain and fever, blood pressure, induction of blood clotting, induction of labor, and the sleep/wake cycle. In most respects the structure of this dimeric enzyme is very similar to other water enzymes, except that it exhibits a highly hydrophobic surface that is built from four α-helices per subunit that are inserted into the lipid bilayer. The use of α-helices to penetrate one side of a lipid bilayer would appear to be a common feature of proteins that interact with one side of a membrane. Interestingly the active site opens to the lipid bilayer via a large tunnel. This provides a thermodynamically acceptable route for arachidonic acid, which is a hydrophobic substrate to enter the enzyme. Prostaglandin synthase is the site of action of nonsteroidal anti-inflammatory drugs such as aspirin and ibuprofen, which act to block the channel that opens to the lipid bilayer and prevent arachidonic acid from being converted to PGH₂.

IX. SUPERFAMILIES AND STRUCTURAL EVOLUTION

The topology of a domain yields information about its evolutionary history. Extensive studies on the sequence variation among a family of similar enzymes found in differing organisms reveal that for a given biological function the protein fold is more conserved than the sequence, except for catalytically vital residues. Remarkably the same protein fold is found in proteins that share no significant sequence similarity. Together these observations suggest that a given fold can accept a wide range of sequences and that evolution preserves the core three-dimensional structure. There are two explanations for these observations. The first explanation suggests that evolution of proteins occurs primarily through point mutations which requires that the evolutionary intermediates must be functional and stable. Clearly this is required if the changes involve an essential enzyme. The same requirement applies to more drastic genetic rearrangements which must also proceed through useful, stable intermediates to survive selective pressure and again this will preserve the protein fold. An alternative explanation is that there are a limited number of stable folds and that enzymes have evolved to reach these conformers. In all likelihood both of these arguments contain elements of truth.

There is no doubt that nature has frequently adapted successful protein architectures to carry out new biological functions. This can be seen clearly in the repeated use of common ligand binding domains (such as the dinucleotide binding motif) that occur repeatedly in proteins that require the same ligand even if it is used in differing chemical reactions. Such use of common building blocks can be rationalized as a consequence of genetic rearrangement. It would appear that some folds, such as the TIM barrel, are particularly well suited for the evolution of new functionalities. The TIM barrel is one of the most abundant enzyme folds and appears to have arisen at least twice during evolution on the basis of the hydrogen bonding pattern observed within the barrel. There are a wide number of enzymes that utilize this fold as the foundation of their active sites which suggests that enzymes may evolve by retooling of existing functional folds.

The enolase superfamily, which contains a variant of the TIM barrel, is a good example of retooling of a functional fold since these enzymes share a common catalytic step of abstraction of the α-proton of a carboxylate anion. This group of enzymes catalyzes a remarkable range of chemical reactions including racemization, β-elimination of water, β-elimination of ammonia, and cycloisomerization. Each enzyme contains similar catalytic bases and acids and each appears to have evolved by reusing a structural framework that facilitates a difficult chemical task. This is consistent with the observation that the protein fold is the component of the structure that changes most slowly during evolution. Fundamentally this is the result of the marginal thermodynamic stability of proteins.

X. QUATERNARY STRUCTURE

Although a substantial number of proteins function as monomers there are many others that exist as multimers. The arrangement of protein subunits in a macromolecular assembly is referred to as its quaternary structure. This aspect of protein structure plays an important role in the stability and regulation of a large number of enzymes, virus assembly, cellular regulation, and motility. Indeed, quaternary structure underlies all aspects of protein–protein interaction.

The magnitude of protein–protein interfaces varies enormously from ~800 Å² to over 4000 Å² per subunit. There is considerable variation in the nature of the interface. Some interfaces are very similar to the hydrophobic core the protein, whereas others contain a substantial polar component and solvent pockets. Thus the stability of a multimeric assembly is only loosely proportional to the surface area buried on formation of the multimeric assembly. As might be expected from considerations of
protein stability, the overall stability of multimeric assemblies is proportional to the hydrophobic surface area buried within the interface.

A large number of enzymes exist as symmetric macromolecular assemblies where they commonly exhibit C and D point group symmetries and contain two-, three-, four-, and six-fold axes of symmetry. In the simplest cases this feature provides additional thermodynamic stability for a protein that would otherwise be rather small. In more complex arrangements the protein–protein interfaces form the active site such that the oligomerization is required for function. Finally, in the most highly evolved enzymes there is communication between the active sites that reside on symmetrically related subunits. This provides the foundation for enzyme regulation as observed in most allosteric enzymes.

A. Allosteric Enzymes

The simplest model for allosteric control was set forward by Monod, Changeux, and Jacob in 1963 and provided the basis for understanding feedback inhibition and cooperative binding of ligands by proteins. In their model it was assumed that an allosteric enzyme (or protein) exists in equilibrium between two symmetric states; inactive and active (T and R states). The transition between these states was assumed to be concerted, that is the symmetry of the macromolecular assembly is conserved. Furthermore the active state has a greater affinity for substrate than the inactive state. From these considerations the activity of the enzyme depends on the position of the equilibrium between the inactive and active states. Thus increasing the substrate concentration drives the equilibrium to the active form and gives rise to a sigmoidal relationship between the initial velocity of the reaction and the substrate concentration. As importantly the position of the equilibrium can be altered by allosteric effectors that preferentially bind to either the inactive or active state. This is the basis of feedback inhibition whereby the product of a biosynthetic pathway inhibits the enzyme that catalyzes the first committed step. This simple model explains many of the properties of allosteric enzymes; however, a more complex model based on sequential binding of substrates to yield multiple conformation states is required to explain the finer details of many enzymes.

The structural basis of allostery has been well developed through the study of enzymes such as aspartate transcarbamoylase and phosphofructokinase. In all enzymes studied thus far several common themes have evolved. First the overall symmetry of the inactive and active states appear to be conserved. Second the transition between these states involves a change in the relationship between the protein subunits that serve to alter the position of the polypeptide chains or cofactors within the regions responsible for ligand binding or catalytic activity. Allosteric effectors serve to either inhibit or enhance these conformational changes. Thus in all cases the active site residues are affected either directly or indirectly by the protein–protein interactions between subunits. Typically a rotation of subunits of 5–15° is sufficient to accomplish allosteric control.

B. Viruses

Viruses represent particularly evolved forms of macromolecular assemblies. Mature virions are encoded by a protective coat that is formed in part by virally encoded proteins. Viruses come in many shapes and sizes, but they all share the property of using multiple copies of coat proteins to protect their genomic material. In many cases the proteins assemble to form a symmetric shell, where the symmetries are either helical as found in tobacco mosaic virus or icosahedral as seen in the spherical viruses. The use of multiple copies of a protein to form a viral coat is enormously efficient from a genomic point of view; however, it introduces several interesting structural problems.

All simple spherical viruses exhibit icosahedral symmetry. This implies that the surface contains 60 equivalent positions or that the shell is built from 60 equivalent protein subunits. Although this is observed for a few very small virus particles, most contain many more than 60 subunits. For example the viral shell of most small plant viruses contains 180 identical protein subunits. This presents a problem of how to arrange 180 subunits on the surface of an icosahedral shell since the subunits cannot experience symmetrically equivalent environments. Fifty years ago it was proposed by Caspar and Klug that viral subunits would be arranged on a hexagonal surface lattice with quasi-equivalent symmetry where the contacts between subunits would be organized to minimize their differences in assembly. That is, if a virus contains 180 subunits on its surface, these would be grouped into three sets of 60 subunits where each group would experience similar interactions compared to the other groups. At first sight this hypothesis accounts for the structure of simple viruses where it has been shown that most of the interactions between the protomers on the surface are essentially identical; however, closer examination of the virus structures that have been determined reveals that in all cases the groups of protomers behave as though they are different proteins by utilizing their domains in different ways to accommodate their structurally unique environments (Fig. 18). Thus, it would appear that the main requirement for a virus coat protein is to have a shape and conformational flexibility to
SEE ALSO THE FOLLOWING ARTICLES

BIOENERGETICS • HYDROGEN BOND • MACROMOLECULES, STRUCTURE • MEMBRANE STRUCTURE • NUCLEAR MAGNETIC RESONANCE (NMR) • PROTEIN FOLDING • PROTEIN SYNTHESIS • X-RAY ANALYSIS

BIBLIOGRAPHY


WEB SITES

Data Bases for Macromolecular Structures

The RCSB (http://www.rcsb.org/pdb/). The most comprehensive archive for coordinates of proteins and nucleic acids (PDB files).

http://www.bmrb.wisc.edu/pages/ A repository for data from NMR spectroscopy on proteins, peptides and nucleic acids

Sites for Structural Analysis


http://mmtsb.scripps.edu/viper/viper.html. Data base for virus structural information

http://www.biochem.ucl.ac.uk/bsm/pdbsum. A site that summarizes structural analysis of PDB files. This contains links to the major structural classification sites.

http://www2.ebi.ac.uk/dali/. The Dali server is a network service for comparing protein structures in 3D. Coordinates of a query protein structure are submitted and Dali compares them against those in the Protein Data Bank. A multiple alignment of structural neighbours is returned.

XI. CONCLUSIONS

Protein structure influences all aspects biological function. Although there is considerable variation in the structural motifs observed in biological macromolecules they are all unified by being built from the same 20 amino acids. The differences are due to the essentially infinite number of protein sequences that may be generated from these building blocks. At a fundamental level, rigorous knowledge of the conformational properties of polypeptide chains should lead to a complete understanding of protein structure and function. As described earlier, great progress has been made toward this goal, but much remains to be done.