3
Protein Structure Introduction

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C^\alpha$</td>
<td>$\alpha$-Carbon</td>
</tr>
<tr>
<td>$\tau$</td>
<td>dihedral angle</td>
</tr>
<tr>
<td>$\phi$</td>
<td>{N–$C^\alpha$} rotation about peptide bond</td>
</tr>
<tr>
<td>$\chi_1$–$\chi_4$</td>
<td>rotamer dihedral angles in amino acid sidechains</td>
</tr>
<tr>
<td>$\psi$</td>
<td>{$C^\alpha$–C}=O rotation about peptide bond</td>
</tr>
<tr>
<td>$\omega$</td>
<td>{$C^\alpha$–C–N}–$C^\alpha_2$ rotation</td>
</tr>
</tbody>
</table>

Life is the mode of existence of proteins, and this mode of existence essentially consists in the constant self-renewal of the chemical constituents of these substances.

Friedrich Engels, 1878 (1820–1895).

3.1 The Machinery of Life

3.1.1 From Tissues to Hormones

The term “protein” originates from the Greek word *proteios*, meaning “primary” or “of first rank”. The name was adapted by Jöns Berzelius in 1838 to emphasize the importance of this class of molecules. Indeed, proteins play
crucial, life-sustaining biological roles, both as constituent molecules and as triggers of physiological processes for all living things. For example, proteins provide the architectural support in muscle tissues, ligaments, tendons, bones, skin, hair, organs, and glands. Their environment-tailored structures make possible the coordinated function (motion, regulation, etc.) in some of these assemblies.

Proteins also provide the fundamental services of transport and storage, such as of oxygen and iron in muscle and blood cells. The first pair of solved protein structures hemoglobin and myoglobin, serve as the crucial oxygen carriers in vertebrates. Hemoglobin is found in red blood cells and is the chief oxygen carrier in the blood (it also transports carbon dioxide and hydrogen ions). Myoglobin is found in muscle cells, where it stores oxygen and facilitates oxygen movement in muscle tissue. The sperm whale depends on myoglobin in its muscle cells for large amounts of oxygen supplies during long underwater journeys.

Proteins further play crucial regulatory roles in many basic processes fundamental to life, such as reaction catalysis (e.g., digestion); immunological and hormonal functions; and the coordination of neuronal activity, cell and bone growth, and cell differentiation.

Given this enormous repertoire, Berzelius could not have coined a better name!

3.1.2 Size and Function Variability

Protein molecules come in a wide range of sizes and have evolved many functions. The major classes of proteins include globular, fibrous, and membrane proteins. Globular proteins are among the most commonly studied group. Newly found ribosomal proteins form a characteristic class of proteins that can be ordered as globular proteins, with disordered extensions.

To suit their environment and function, fibrous proteins (e.g., the collagen molecule in skin and bones), which are generally insoluble in aqueous environments, are extended in shape, whereas globular proteins tend to be compact. Collagen is a left-handed helix with a quaternary structure made of collagen fibrils aggregated in a parallel superhelical arrangement. See [680] for the crystal structure of a collagen-like peptide with a biologically relevant sequence (also shown in Figure 3.9) and summary of collagen structures elucidated to date. The globular protein myoglobin (see Figure 3.12) is highly compact, organized as 75% α-helices. Similarly, hemoglobin is a tetramer composed of four polypeptide chains held by noncovalent interactions; each subunit of hemoglobin in humans is very similar to myoglobin. Both proteins bind oxygen molecules through a central heme group.

There certainly are some very large proteins such as the muscle protein titin of about 27000 amino acids (and mass of 3000 kDa), but the average protein contains several hundred residues. The size of polypeptides can be determined from gel electrophoresis experiments: the rate of migration of the molecule is inversely proportional to the logarithm of its length. The mass of a polypeptide or protein can be estimated from mobility-to-mass relationships established for reference proteins and by mass spectrometry measurements. Equilibrium ultracentrifugation
Figure 3.1. (left) The general formula for an amino acid, and (right) the spatial tetrahedral arrangement of an L-amino acid. The mirror image, a D-amino acid, is rare in proteins in Nature, if it exists.

...is another favored technique for determining various macromolecular features, including molecular weight, on the basis of transport properties.

### 3.1.3 Chapter Overview

This chapter introduces the bare basics in protein structure: the amino acid building blocks, primary sequence variations, and the framework for describing conformational flexibility in polypeptides. Included also is an introduction to the more advanced topic of sequence similarity and relation to structure, in the section on variations in protein sequences. *Students are encouraged to return to this subsection after reading Chapters 3 and 4.* Chapter 4 continues to describe secondary, supersecondary, and tertiary structural elements in proteins, as well as protein classification.

The protein treatment in these chapters is brief in comparison to the minitutorial on nucleic acids. Readers should consult the many excellent texts on protein structure (see books listed in Appendix C), like that by Branden and Tooze, and review introductory chapters in biochemistry texts like Stryer’s. The 1999 text by Fersht [394], in particular, is a comprehensive description of the state-of-the-art in protein structure, and also reviews recent advances and insights from theoretical approaches.

---

**Box 3.1: Water Structure**

Water — that deceptively simple molecule composed of two hydrogens and one oxygen — displays highly unusual and complex properties that are far from fully understood [788]. Perhaps because of those properties — like the contraction of ice when it melts, large heat of vaporization, and large specific heat — water is the best of all solvents and a fundamental substance to sustain life. Solvent organization and reorganization are crucial to the stability of proteins, nucleic acids, saccharides, and other molecular systems and...
profoundly [384]. The energetic and kinetic aspects of water structure are difficult to pinpoint by experiment and simulation because of the range of timescales associated with water motions, from the fast perturbations of order 0.1 ps to the slow proton exchanges of millisecond order.

Important to the understanding of solvation structure and dynamics in the vicinity of macromolecules is the tendency of water to form hydrogen bonds [1018] (see also Box 3.2 for a definition of a hydrogen bond). In ice, the ordered crystal structure of water molecules, each oxygen is surrounded by a tetrahedron of four other oxygen atoms, with one hydrogen between each oxygen pair. In liquid water, many water molecules are engaged in such a hydrogen-bonded network, but the network is highly dynamic, with hydrogen-bonded partners changing rapidly.

This local organization of liquid water can easily be observed from experimental and computed radial distribution functions (e.g., O–O and O–H distances), which reflect the degree of occupancy of neighbors from a central oxygen or hydrogen molecule. Thus, for example, the highest peak in the O–O radial distribution function at a distance of about 2.9 Å corresponds to the first solvation shell, in which the four near-neighbor oxygens of the central oxygen molecule can be found at room temperature.

Figure 3.2 illustrates the structure of water clusters as computed by minimizing the potential energy composed of bond length, bond angle, and intermolecular Coulomb and van der Waals terms (see Chapter 9 for energy terms discussion). Such hydrogen bonds form ubiquitously in the environment of biomolecules. Water molecules penetrate into the grooves of nucleic acid helices, aggregate around hydrophilic, or water-soluble, segments of proteins (which cluster at the protein surface) and stabilize solute conformations through various hydrogen bonds and bridging arrangements. The dynamic nature of both water structure and biomolecules gives rise to the concept of hydration shells; see Chapter 6 in the context of DNA. That is, the solvent structure around the solute is multilayered, with the first hydration shell associated with water molecules in direct contact with the solute and the outermost layer as the bulk solvent.

---

**Box 3.2: Hydrogen Bonds**

A hydrogen bond is an attractive, weak electrostatic (noncovalent) bond [1018]. It forms when a hydrogen atom covalently binds to an electronegative atom and is electrostatically attracted to another (typically electronegative) atom. The atom to which the hydrogen atom (H) is covalently bound is considered the hydrogen donor (D), and the other atom is the hydrogen acceptor (A). Thus, the hydrogen bond is stabilized by the Coulombic attraction between the partial negative charge of the A atom and the partial positive charge of H. See Figure 3.2 for examples in water clusters.

In biological polymers, the donor and acceptor atoms are either nitrogens or oxygens. For example, in protein helices and sheets, the D–H···A sequence is N–H···O=C. In the nucleic-acid base pair of adenine–thymine, the two D–H···A sequences are N–H···O and N–H··· (see Nucleic Acid chapters for details). Non-classical, weaker hydrogen
Figure 3.2. Structures of water clusters of 2, 4, 8, 16, and 125 molecules as minimized with the CHARMM force field from initial coordinates computed in [1120]. Two geometries are shown for both the 8 and 16-molecule clusters; the lower energy structures (by roughly 4 and 6%, respectively) are associated with the more compact, cube-like shapes (left side in both cases). The tetrahedral structure of water is apparent in the inner molecules of the larger systems, where each molecule is hydrogen bonded to four others.
bonds have been noted in biological systems (e.g., protein/DNA complexes), involving a carbon instead of one electronegative atom: C–H···O [820].

The strength of a hydrogen bond can be characterized by two geometric quantities which govern the hydrogen bond energy: colinearity of the D–H···A atoms, and optimal H···A (or D···A) distance. The ideal, strongest hydrogen bond often has its three atoms colinear. The strength of a hydrogen bond is several kilocalories per mole, compared to about 0.6 kcal/mol for thermal energy at room temperature, but the exact value remains uncertain (e.g., [319]). However, the formation of a network of hydrogen bonds in macromolecular systems leads to a cooperative effect that enhances stability considerably [1018].

3.2 The Amino Acid Building Blocks

Proteins and polypeptides are composed of linked amino acids. That amino acid composition of the polymer is known as the primary structure or sequence for short.

3.2.1 Basic Cα Unit

Each amino acid consists of a central tetrahedral carbon known as the alpha (α) carbon (Cα) attached to four units: a hydrogen atom, a protonated amino group (NH₃⁺), a dissociated carboxyl group (COO⁻), and a distinguishing sidechain, or R group (see Figure 3.1).

This dipolar or zwitterionic form of the amino acid (COO⁻ and NH₃⁺) is typical for neutral pH (pH of 7). The un-ionized form of an amino acid corresponds to COOH and NH₂ end groups. Different combinations involving ionized/un-ionized forms for each of the side groups can occur for the amino acid depending on the pH of the solution.

![Figure 3.3. Formation of a dipeptide by joining two amino acids.](image-url)
3.2. The Amino Acid Building Blocks

The tetrahedral arrangement about Cα makes possible two mirror images for the molecule. Only the L-isomer (“left-handed” from the Latin word *levo*) is a constituent of proteins on earth (see Figure 3.1). This asymmetry is not presently understood, but one explanation is that this imbalance is related to an asymmetry in elementary particles.

### 3.2.2 Essential and Nonessential Amino Acids

There are 20 naturally-occurring amino acids. Among them, humans can synthesize about a dozen. The remaining 9 amino acids must be ingested through our diet; these are termed *essential amino acids*. They are histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine.

Meat eaters need not be too concerned about a balanced diet of those nutrients, since animal flesh is a complete source of the essential amino acids. In contrast, vegetarians, particularly vegans — who omit all animal products like eggs and dairy in addition to meat, poultry, and fish — must perform a delicate balancing act to ensure that their bodies can synthesize all the basic proteins essential to good health. See Box 3.3 for a discussion of essential amino acids and balanced diets.

---

**Box 3.3: Protein Chemistry and Vegetarian Diets**

Vegetarians must take care to combine foods from three basic groups which are complementary with respect to their supply of these amino acids: (a) rice and grains like oats, wheat, corn, cereals, and breads; (b) legumes and soy-products; and (c) nut products (cashews, almonds, and various nut butters). Peanuts are technically members of the legume family rather than nuts.

Notable vegetarian combinations are: rice or grains (low in or lacking isoleucine, lysine, and threonine) with beans (rich in isoleucine and leucine and, in the case of lima beans, also lysine); cereals with leafy vegetables; corn, wheat, or rye (low in or lacking tryptophan and lysine) with soy protein/soybeans (rich in isoleucine, tryptophan, lysine, methionine, and valine); corn with nuts or seeds (rich in methionine, isoleucine, and leucine); bread/wheat with peanut butter (rich in valine and tryptophan); and potatoes (limited methionine and leucine) with onions, garlic, lentils and egg or fish (if permitted), all of which are rich in methionine.

A classic Native-American dish of acorn squash stuffed with wild rice, quinoa, and black beans is a superior mixture of nutrients. The plant quinoa, prepared like a grain, is actually a fruit and moreover a complete protein (rich in lysine and other amino acids), as well as rich in vitamins E and B, fiber, and the minerals calcium, phosphorus, and iron. Nuts also contain several vitamins and minerals that protect against heart disease, like folate, and calcium, magnesium, and potassium, which also protect against high blood pressure.

---

1. At least two nonstandard by naturally occurring amino acids in certain Archaea and eubacteria are known, pyrrolysine and selenocysteine [68, 511].
Contrary to an existing myth, such food group combinations need not be eaten at the same meal to guarantee a complete source of the essential amino acids; a daily approach suffices. Given all these food sources, vegetarians — even vegans who carefully comply — will not be deficient in protein. However, nutrients that present a challenge to vegans and lacto-ovo vegetarians are vitamin B_{12} (deficiencies of which can cause nerve damage), found in fortified cereals, and zinc (needed for protein synthesis, healing wounds, and immunity), available in fortified cereals, soy-based foods, and dairy products.

*Given the intimate relationship between protein chemistry and good nutrition, readers of this text should be healthy as well as smart!*

Interestingly, the requirements for vegetarian diets are now of crucial interest to NASA researchers: future astronauts who will spend extended periods of time in space stations (on Mars, Jupiter, or the Moon) will have to depend on hydroponically-grown plant crops for nearly all their protein requirements, as well as vitamins, minerals, and fiber. Research is now in progress on how best to select a limited set of plants that can adapt to growing in nutrient-enriched water (rather than soil) and in small spaces. At the same time, this selection must meet the basic dietary requirements of space-station scientists, as well as satisfy their culinary taste and demand for variety [173].

![Figure 3.4. The repeating formula for a polypeptide.](image)

Aspartame

![Figure 3.5. The dipeptide aspartame.](image)
3.2.3 Linking Amino Acids

A polypeptide is formed when amino acids join together. Namely, the carboxyl carbon of one amino acid joins the amino nitrogen of another amino acid to form the peptide (C–N) bond with the release of one water molecule (Figure 3.3). The general repeating formula for a polypeptide is shown in Figure 3.4. When the amino acid residue is proline, its Cα is linked to the nitrogen of the peptide backbone through the proline ring.

A model of aspartame, a dipeptide of aspartic acid and phenylalanine, is shown in Figure 3.5. It was discovered accidentally in 1965 by a ‘careless’ chemist who licked his fingers during his laboratory work. To his surprise, a substance 100–200 times sweeter than sucrose was discovered. Because it is a kind of protein, aspartame is metabolized in the body like proteins and is a source of amino acids. *(This should not, however, be taken as an endorsement for diet soft drinks as a source of nutrients!)*

The synthesis of polypeptides and proteins occurs in a cellular structure, the ribosome, *in vivo*. Synthesis *in vitro* is facile for 100–150 residues but much more involved for longer chains.

3.2.4 The Amino Acid Repertoire: From Flexible Glycine to Rigid Proline

The chemical formulas of the twenty L-amino acids are shown in Figure 3.6, with the corresponding space-filling models shown in Figure 3.7. The commonly used three-letter abbreviation for each acid is illustrated, as well as a grouping into amino acid subfamilies. A one-letter mnemonic is also used to identify sequences of amino acids, as shown in Table 3.1.

A broader classification than indicated in the figures consists of the following three groups:

- **NPo**: amino acids with strictly nonpolar (hydrophobic, or water insoluble) side chains: Ala, Val, Leu, Ile, Phe, Pro, Met, Gly, Trp, Tyr;
- **CPo**: amino acids with charged polar residues: Asp, Glu, His, Lys, Arg; and
- **UPo**: amino acids with uncharged polar side chains: Ser, Thr, Cys, Asn, Gln.

Each amino acid has a unique combination of properties — size, polarity, cyclic constituents, sulfur constituents, etc. — that critically affects the noncovalent and covalent (i.e., disulfide bonds) interactions that form and stabilize protein three-dimensional (3D) architecture. These interactions originate from electrostatic, van der Waals, hydrophobic, and hydrogen bonding forces. These properties are described in turn for these amino acid classes.
### Protein Structure Introduction

**Aliphatic Side Chains (NPo)**
- **Glycine (Gly)**
- **Valine (Val)**
- **Leucine (Leu)**
- **Isoleucine (Ile)**

**Aliphatic Hydroxyl Side Chains (UPo)**
- **Serine (Ser)**
- **Threonine (Thr)**

**Secondary Amino Group (NPo)**
- **Proline (Pro)**

**Acidic Side Chains and Their Amide Derivatives (CPo - Asp, Glu; UPo - Asn, Gln)**
- **Aspartic Acid (Asp)**
- **Glutamic Acid (Glu)**
- **Asparagine (Asn)**
- **Glutamine (Gln)**

**Sulfur-Containing Side Chains (NPo)**
- **Methionine (Met)**
- **Cysteine (Cys)**

**Basic Side Chains (CPo)**
- **Lysine (Lys)**
- **Arginine (Arg)**
- **Histidine (His)**

**Aromatic Side Chains (NPo but potential for polarity)**
- **Phenylalanine (Phe)**
- **Tyrosine (Tyr)**
- **Tryptophan (Trp)**

---

Figure 3.6. The chemical formulas of the 20 natural amino acids as found in neutral pH (pH of 7). The acronyms NPo, UPo, CPo denote, respectively, nonpolar, uncharged polar, and charged polar amino acids.
Aliphatic R: Gly, Ala, Val, Leu, Ile

Glycine, alanine, valine, leucine, and isoleucine can be classified as nonpolar. Glycine, the simplest of the amino acids, is first in the aliphatic-sidechain subgroup. Each member in this family has a sidechain (R) which increases in...
bulk and branching design. Glycine is most flexible and hence an important constituent of proteins. For example, glycine is a major component of the \( \alpha \)-helix of the protein \( \alpha \)-keratin, which makes up hair and wool, as well as the \( \beta \)-sheet of the polypeptide \( \beta \)-keratin, which is silk (see Figure 3.9). Since the increasing aliphatic substitutions in this family increases the bulkiness of the amino acid, the overall conformational flexibility correspondingly decreases within a polypeptide. However, the conformational variability of each of these amino acids increases due to the *rotameric* variations of the amino acid (roughly, different 3D arrangements about central bonds within the sidechain — see Section 3.4).

**Rigid Proline**

Proline is a nonpolar amino acid as well. In contrast to glycine residues, which allow a great deal of conformational flexibility about the backbone (i.e., wide range of sterically-permissible rotations \( \phi \) and \( \psi \) about the peptide bond — see Section 3.4), flexibility in proline residues is largely limited, due to the cyclic nature of its sidechain.

**Aliphatic Hydroxyl R: Ser, Thr**

Serine and threonine contain aliphatic hydroxyl groups and are considered uncharged polar, capable of forming hydrogen bonds (see Box 3.2).

**Acidic R and Amide Derivatives: Asn, Gln, Asp, Glu**

Similarly, asparagine and glutamine possess amide groups and are also considered uncharged polar with potential for hydrogen bond formation. Their acidic analogs, aspartic acid and glutamic acid, are negatively charged (intrinsic pH of around 4) and thus considered charged polar, but the polar end of their sidechains is separated from \( C^\alpha \) by hydrophobic \( CH_2 \) groups.

**Basic R: Lys, Arg, His**

Lysine, arginine, and histidine have basic sidechains and are thus in the charged polar category of amino acids. Lysine and arginine — the longest amino acids — are positively charged at physiological concentrations (that is, sidechain pH of 10–12), whereas histidine’s charge can be both positive or negative depending on its environment. This duality in histidine stems from its imidazole ring, which is in the physiological range of pH. For this reason, histidine residues serve as good metal binders and are often found in the active sites of proteins.

**Aromatic R: Phe, Tyr, Trp**

The amino acids with aromatic sidechains — phenylalanine, tyrosine, and tryptophan — have significant potential for electrostatic interactions due to an electron deficit in the ring hydrogen atoms. Phenylalanine is highly hydrophobic while the
other two can be considered mildly hydrophobic, since their aromaticity is juxta-
posed with polar properties (hydroxyl group of tyrosine and indole-ring nitrogen of
tryptophan). The aromatic rings of this amino acid family also have potential for electron transfer. They can all be classified as nonpolar, though the mild hydrophobicity of tyrosine often warrants its classification as an uncharged polar amino acid.

Sulfur-Containing R: Met, Cys

Finally, nonpolar cysteine and methionine contain sulfur in their sidechains and are thus hydrophobic. Cysteine, in particular, is very reactive and binds to heavy metals. It has an important role in protein conformations through its unique ability to form disulfide bonds between two cysteine residues. Disulfide bonds are covalent but reversible and are thought to be important in many cases by directing a protein to its native structure and maintaining this functionally-important state.

Figure 3.8. Amino acid frequencies as computed from 45,137 proteins collected from 15 taxa representing the three kingdoms of life (Bacteria, Archaea, and Eukaryota) [615]. See Table 3.1 for the frequencies and key to the one-letter amino acid abbreviations.

3.3 Sequence Variations in Proteins

From the constituent library of twenty natural amino acids, $20^N$ sequence combina-
tions for an $N$-residue peptide are possible, an enormous number when $N$ is several hundred. However, natural evolution has favored certain sequences more than others. Sequence similarity is an important factor in indicating common evolu-
tionary ancestry of proteins, as discussed below. It is therefore widely used as a tool for classifying proteins into families, as well as for relating sequence to structure and predicting structure from sequence (homology modeling [16,79], as introduced in Chapter 2).
3.3.1  **Globular Proteins**

In most proteins, the twenty amino acids occur at roughly similar frequencies. Notable exceptions occur for certain amino acids like methionine, which is frequently found at the N-terminus of the peptide since it serves as the amino acid initiator of synthesis, or special groups of proteins, such as membrane or fibrous proteins.

Table 3.1 shows the frequency of occurrence of amino acid residues in the PDB40 dataset of 971 domains of unrelated proteins with a sequence identity of 40% or less [959], and Figure 3.8 displays the data as histograms. We see that nonpolar Ala and Leu (boldfaced entries in the table) have the highest percentages (above 8%) within the representative protein database. The lowest frequencies (4% and below) occur for Trp (aromatic sidechain), Cys (sulfur-containing sidechain), His and Met (the other sulfur-containing sidechain), Tyr and Phe (aromatic sidechains also), and Gln.

Table 3.1. Amino acid frequencies in proteins based on the data of [615] which analyzed 45,137 proteins from 15 taxa. Bold and italics types are used, respectively, for the highest (>8%) and lowest (≤ 2.5%) frequencies.

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Freq. [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Alanine</strong> (Ala, A)</td>
<td><strong>8.1</strong></td>
</tr>
<tr>
<td>Arginine (Arg, R)</td>
<td>5.1</td>
</tr>
<tr>
<td>Asparagine (Asp, D)</td>
<td>5.2</td>
</tr>
<tr>
<td>Aspartic acid (Asn, N)</td>
<td>4.0</td>
</tr>
<tr>
<td>Cysteine (Cys, C)</td>
<td>1.2</td>
</tr>
<tr>
<td>Glutamine (Gln, Q)</td>
<td>3.8</td>
</tr>
<tr>
<td>Glutamic acid (Glu, E)</td>
<td>6.5</td>
</tr>
<tr>
<td>Glycine (Gly, G)</td>
<td>7.2</td>
</tr>
<tr>
<td>Histidine (His, H)</td>
<td>2.2</td>
</tr>
<tr>
<td>Isoleucine (Ile, I)</td>
<td>6.8</td>
</tr>
<tr>
<td><strong>Leucine</strong> (Leu, L)</td>
<td><strong>10.3</strong></td>
</tr>
<tr>
<td>Lysine (Lys, K)</td>
<td>5.9</td>
</tr>
<tr>
<td><strong>Methionine</strong> (Met, M)</td>
<td>2.5</td>
</tr>
<tr>
<td>Phenylalanine (Phe, F)</td>
<td>4.2</td>
</tr>
<tr>
<td>Proline (Pro, P)</td>
<td>4.3</td>
</tr>
<tr>
<td>Serine (Ser, S)</td>
<td>6.2</td>
</tr>
<tr>
<td>Threonine (Thr, T)</td>
<td>5.1</td>
</tr>
<tr>
<td>Tryptophan (Trp, W)</td>
<td>1.1</td>
</tr>
<tr>
<td>Tyrosine (Tyr, Y)</td>
<td>3.2</td>
</tr>
<tr>
<td>Valine (Val, V)</td>
<td>6.9</td>
</tr>
</tbody>
</table>

3.3.2  **Membrane and Fibrous Proteins**

Membrane proteins are embedded in a dynamic lipid bilayer environment, where mobility is more restricted. They therefore have more hydrophobic residues than
globular proteins, which favor polar groups on the exterior surface. Since membrane proteins are particularly difficult to crystallize, simulation work is especially important in this area to understand their detailed function.

Fibrous, or structural, proteins tend to have repetitive sequences. The triple-stranded collagen helix, for example, is composed of repeating triplets which include glycine as the first residue and often proline as one or both of the other residues of the triplet. A model of collagen is shown in Figure 3.9.

Since collagen is needed to rebuild joint cartilage, there are important practical applications to skin and bone ailments. For example, a gelatin-containing (glucosamine and calcium-enriched) powdered drink mix called Knox NutraJoint is being touted as a dietary supplement that helps maintain healthy joints and bones. (‘Juice Your Joints’ touts an ad featuring an athletic sexagenarian water skier). Gelatin is rich in two amino acids, glycine and proline, that make up collagen. Even though our bodies make these two amino acids, manufacturers claim that this gelatin-containing supplement may be helpful in decreasing the progression of osteoarthritis, a condition caused by cartilage deterioration.

Figure 3.9. Models of the fibrous proteins collagen (triple helix) and silk, along with a crystallographically-determined collagen-like peptide (Hyp denotes hydroxyproline).

Another use of collagen is in a skin product used to heal wounds such as from venous skin ulcers, burns, and skin surgery. In May 1998 the FDA approved Apilgraf, a product for treating venous skin ulcers made of human skin cells mixed with collagen cells from cattle.
Silk is another example of a fibrous protein with a repetitive sequence. The product of many insects and spiders, silk is the polypeptide β-keratin composed largely of glycine, alanine, and serine residues, with smaller amounts of other amino acids such as glutamine, tyrosine, leucine, valine, and proline. The softness, flexibility, and high tensile strength of silk stems from its unique arrangement of loose hydrogen bonding networks in the form of β-sheets connected by β-turns, a mixture of both highly-ordered and less densely-packed regions. Figure 3.9 shows a model of the repetitive β-sheet network of silk (without connecting regions).

3.3.3 Emerging Patterns from Genome Databases

As genome sequencing projects are completed, interesting findings on enzyme sequences also emerge. For example, the genome of the tuberculosis bacterium (completed in 1998 by the Wellcome Trust Genome Campus of the Sanger Institute in collaboration with the Institut Pasteur in Paris) revealed surprisingly that, unlike other bacteria, repetitive gene families of glycine-rich proteins exist in *M. tuberculosis*; these approximately 10% of the enzyme-coding sequences are associated with gene families involved in anaerobic respiratory functions.

3.3.4 Sequence Similarity

Sequence Similarity Generally Implies Structure Similarity

As mentioned above, sequence similarity generally implies structural, functional, and evolutionary commonality. Thus, for example, if we were to scan the Protein Databank (PDB) randomly and find two proteins with low sequence identity (say less than 20%), we could reasonably propose that they also have little structural similarity. Such an example is shown in Figure 3.12 for the cytochrome/barstar pair. Similarly, large sequence similarity generally implies structural similarity (see introduction in 2.1.2 of Chapter 2).

In general, small mutations (e.g., single amino acid substitutions) are well tolerated by the native structure, even when they occur at critical regions of secondary structure. The small protein Rop (Repressor of primer), which controls the mechanism of plasmid replication, provides an interesting subject to both this sequence-implies-structure paradigm, and to exceptions to this rule (discussed below).

Rop is a dimer, with each monomer consisting of two antiparallel α-helices connected by a short turn; it dimerizes to form a four-helix bundle as active form, as shown in Figure 3.10. (Fold details and motifs are discussed in the next chapter). Recall that Rop was used as the basis for solving Paracelsus challenge (Chapter 2) because the α-helix motif was thought to be quite stable.

The high stability of Rop emerged surprisingly from experiments of Castagnoli *et al.* [205]. When these researchers deleted just a few residues in a key turn region that produces the overall bundle fold in the native Rop structure, they expected one long contiguous helix to form. Instead, their tinkering produced a small variation
Figure 3.10. The protein Repressor of Primer (63 residues per monomer) provides interesting examples of the paradigm of structure inference by sequence similarity: the four-helix bundle motif of the wildtype (a) can be both structurally stable, i.e., resistant to mutations — as shown by the two variants in (b) and (c) — or structurally fragile and highly sensitive to mutations, caused by proline substitution at the turn region — as shown in (d), a mutant with an entirely different topology [462]. In (b), two Ala residues were inserted at both sides of the amino acid Asp in the loop region. In (c), the Asp residue connecting the two α-helices of each monomer was mutated to Gly, and Met1 was changed to Gly [205].
of the original bundle motif. Apparently, the four-helix bundle motif is so stable that a new turn was formed from residues that used to be part of the \( \alpha \)-helix backbone! Thus, the original bundle motif, though slightly smaller, was maintained in the mutants. This is seen in Figure 3.10, which displays the wildtype enzyme structure (a) and those of two mutants in the above cited study (b and c).

Though this experiment supports the general notion that protein structures are remarkably stable to tinkering (mutations), we emphasize that functional properties of proteins are fragile and quite sensitive to sequence changes.

Exceptions Exist

There are many exceptions, however, to this simple sequence/structure/function relationship.

Namely, examples exist where despite large sequence similarity there is small structural and functional similarity. A classic example of this relationship is the disease sickle-cell anemia, where a minute substitution in sequence leads to altered function with devastating consequences. This abnormality results from the replacement of the highly-polar glutamate residue in hemoglobin by the nonpolar amino acid valine. This key substitution at the surface of the protein leads to an entirely different quaternary structure for this multidomain red-blood pigment protein. This is because the markedly altered structure affects the solubility of oxygenated hemoglobin and leads to a clumping of the deoxygenated form of the molecule (HbS instead of HbA).

Conversely, examples exist where despite small sequence similarity there is large structural, and even functional and evolutionary similarity. A classic example for this relationship is the myoglobin/hemoglobin pair of proteins (see Figure 3.12). These proteins only share 20% of the sequence. However, as oxygen-carrying molecules, they share structural, functional, and evolutionary similarity. Proteins in the calmodulin family are also known to display a great deal of structural variability for similar sequences \([664]\) (see Figure 3.11).

More generally, changes in 3D architecture (despite a nontrivial degree of sequence similarity) can result from a variety of factors, as follows.

- Mutations in critical regions of the proteins, such as active sites and ligand binding sites, can change 3D structures dramatically. Such an example is shown for the pair of immunoglobulins in Figure 3.12.

- Mutations in regions that connect two secondary-structural elements can also be responsible for structural divergence, as in the helix-loop-helix motif of the EF-hand family, and the connecting loops in helix bundles.

Figure 3.11 illustrates this principle for the two EF-hand calcium-binding proteins calmodulin and sarcoplasmic calcium-binding protein: one is overall extended in shape while the other is more compact \([979]\).

Helix bundles are sensitive to mutations in loop or turn regions that connect different helices, to the extent that a single amino acid substitution (alanine
to proline) can change the topology of a homodimeric 4–helical bundle protein from the canonical left-handed all-antiparallel form to a right-handed mixed parallel and antiparallel bundle [462]. Figure 3.10(d) shows this different resulting topology of the **Rop four-helix bundle** subject to the single mutation Ala31→Pro at the turn region.

- **Structural variations** can be observed in the same system determined at different *environmental conditions* such as solvent or crystal packing. The same **T4-lysozyme** mutants in Figure 3.12 (100% sequence similarity) display intriguing mobility, adopting 5 different crystal conformations [374] due to a hinge bending motion.

- **Multidomain proteins** can adapt quaternary structures that depend sensitively on the number of subunits and/or on the sequence.

---

**Figure 3.11.** Structural variability despite large similarity in the protein secondary-structural elements is illustrated for two calcium-binding proteins — **calmodulin** (148 residues) and **sarcoplasmic calcium binding protein** (174 residues) [979] — due to different overall 3D arrangement of the shared motifs. Though sharing only 30% of the sequence, both proteins are made of two repeating units, each consisting of two EF hand motifs. Each hand motif contains two helical regions surrounding a calcium binding loop (crystal-bound calcium atoms are rendered as large spheres; only three are bound to 2scp).
Figure 3.12. Various examples of sequence/structure relationships in proteins: (a) Low sequence similarity (3% for alignment of 72% of the residues) generally implies low structure similarity (cytochrome C6 versus barstar). Still, exceptions are found. For example, in (b), despite low (12%) sequence similarity, there is large structure and function similarity (hemoglobin and myoglobin); conversely, despite high sequence similarity, there can be structural diversity, due to (c) hinge bending in two lysozyme mutants (Met → Ile in residue 6) or (d) different orientation of one of the two subunits in two immunoglobulins. The lysozyme mutant displays 5 different crystal conformations, one similar to the wild-type (shown in blue) and others overall very similar except for a different hinge-bending angle (see defining arrows); the form with largest bend (32°) is shown in yellow. The two immunoglobulins differ markedly in tertiary organization due mainly to differences in the linker domain between the A and B subunits of each protein.
Figure 3.13. Gauche (g) and trans (t) dihedral-angle orientations for n-butane: (left) classification wheel; (middle) simple Newman projections that illustrate the three favored orientations of the two end methyl groups about the central C–C bond (perpendicular to the plane of the paper); and (right) the trans conformation, which has the least steric clashes.

3.4 Protein Conformation Framework

3.4.1 The Flexible $\phi$ and $\psi$ and Rigid $\omega$ Dihedral Angles

Polypeptides can have a wide variety of conformations, i.e., 3D structures differing only in rotational orientations about covalent bonds. This type of rotational flexibility is characterized by a dihedral angle, which measures the relative orientation of four linked atoms in a molecule, $i-j-k-l$. A dihedral angle for a 4-atom sequence that is not necessarily covalently bonded can also be used for special terms in the potential energy function; see Chapter 9 for examples. See Box 3.4, Figure 3.14, and the equations in Appendix D, under the Addendum to Assignment 8 for a definition of a dihedral angle.

---

**Box 3.4: Dihedral Angle**

The dihedral angle $\tau_{ijkl}$ defined for a sequence of linked atoms $i-j-k-l$ (Figure 3.14) is the angle between the normal to the plane of atoms $i-j-k$ and the normal to the plane of atoms $j-k-l$. The sign of $\tau_{ijkl}$ is determined by the triple product $(a \times b) \cdot c$, where $a$, $b$ and $c$ are the interatomic distance vectors for atoms $i \rightarrow j$, $j \rightarrow k$ and $k \rightarrow l$, respectively.

Strictly defined, the related torsion angle, $\tilde{\tau}$ is the angle between the two planes defined by $i-j-k$ and $j-k-l$. Thus, $\tau + \tilde{\tau} = \pi$ ($180^\circ$). However, the terms torsion and dihedral angle are often used interchangeably. We will often use dihedral angle to refer to the numerical value of the angle, and torsion angle or torsional potential when we discuss general properties of these rotations.

---

$^2$see Chapter 8, Subsection 8.4.1, for the related definition of configuration.
When the dihedral angle is 0°, the four atoms \( i-j-k-l \) are coplanar and atoms \( i \) and \( l \) coincide in their projections onto the plane normal to the \( j-k \) bond; this orientation is defined as \( \text{cis} \) or \( \text{syn} \). When the dihedral angle is 180°, the atoms are coplanar but atoms \( i \) and \( l \) lie opposite one another in the projection onto the plane normal to the \( j-k \) bond; such an orientation is defined as \( \text{trans} \) or \( \text{anti} \). More generally, angular regions convenient to describe protein and nucleic acid conformations are the following: \( \text{cis} \) (\( \approx 0° \)), \( \text{trans} \) (\( \approx 180° \)) and \( \pm \text{gauche} \) (\( \approx \pm 60° \)). Another common terminology is: \( \text{syn} \) (\( \approx 0° \)), \( \text{anti} \) (\( \approx 180° \)), \( \pm \text{synclinal} \) (\( \approx \pm 60° \)), and \( \pm \text{anticlinal} \) (\( \approx \pm 120° \)). See Figure 3.13 for a simple illustration for \( n \)-butane.

While the peptide group (Figure 3.3) is relatively rigid — it has 40% double-bond character — there is a great deal of flexibility about each of the single bonds along the backbone, \{N–C\( ^\alpha \)\} and \{C\( ^\alpha \)–C\}=O. The two dihedral angles \( \phi \) and \( \psi \) are used to define rotations about the bond between the nitrogen and C\( ^\alpha \) of the mainchain and between C\( ^\alpha \) and the carbonyl carbon, respectively (Figure 3.15).

The dihedral angle \( \omega \) defines the rotation about the peptide bond, namely for the atomic sequence C\( ^\alpha _1 \)–{C–N}–C\( ^\alpha _2 \), where C\( _1 \) and C\( _2 \) are the \( \alpha \)-carbons of two adjacent amino acids. Because of the partial double-bond character of the peptide bond and the steric interactions between adjacent sidechains, \( \omega \) is typically in the \( \text{trans} \) configuration: \( \omega = 180° \).\(^3\) In this orientation, all four atoms lie in the same plane, with the distance between C\( ^\alpha _1 \) and C\( ^\alpha _2 \) as large as possible (see Figure 3.13 for a definition of various dihedral angle orientations).

\[^3\]Non-trivial deviations from planar peptide bonds can be shown by theory and experiment (e.g., as reviewed in [348]). A statistical survey of peptide and protein databases verified that the distribution of rotation angles (or energies associated with peptide bond rotations) follows Boltzmann statistics [799].
3.4.2 Rotameric Structures

Besides the \{\phi, \psi\} flexibility associated with the two backbone bonds involving \(C^\alpha\), multiple conformations are possible for 18 of the 20 amino acids when the sidechain geometries differ (excluded are glycine and alanine). Rotameric structures of amino acids (and hence proteins) are those that have the same \{\phi, \psi\} angles but differ in the sidechain conformations. The dihedral angles used to define sidechain rotations are denoted by \(\chi\), with subscripts used as needed (see Figure 3.17).

For example, in lysine, whose sidechain has four carbons (see Figure 3.16), dihedral angles \(\chi_1\) through \(\chi_4\) denote the rotations about bonds \(C^\alpha-C_1\), \(C_1-C_2\), \(C_2-C_3\), and \(C_3-C_4\), respectively (see also Figure 3.17 for other amino acids). Rotameric structures for polypeptides and proteins depend on the environment of the polymer and on the secondary and tertiary structures.

3.4.3 Ramachandran Plots

The feasible combinations of the \(\phi\) and \(\psi\) angles are limited due to steric hindrance. That is, only certain combinations are typically observed, with some dependence on residue size and shape. Glycine is unique in its flexibility — it is therefore a good agent for turns in polypeptides and proteins — but other residues exhibit a highly limited range of sterically-permissible \(\phi\) and \(\psi\) combinations. In fact, only roughly one tenth the area of the \{\phi, \psi\} space is generally observed for polypeptides and proteins. Among the first to note this limitation were
Figure 3.17. Rotameric notation used for 18 of the 20 amino acids is illustrated using different colors for $\chi_1$–$\chi_4$, as shown in the top right key.
G.N. Ramachandran⁴ and coworkers in 1963, after which \textit{Ramachandran plots} are called. Around the same time, John Schellman and coworkers were working independently along the same lines of mapping the energetically favorable and excluded regions for protein conformations [1098].

These diagrams in the \{φ, ψ\} space, as shown in Figure 3.18, are used to describe this \{φ, ψ\} flexibility (actually inflexibility) in polypeptides and proteins. See also Figure 3.19 for a comparative view of Ramachandran plots derived from the moderate-resolution X-ray structures shown in Figure 3.18 versus high-resolution X-ray as well as NMR-derived structures.

Often, Ramachandran diagrams are presented by plotting the backbone dihedral angles of all nonterminal residues in a protein for a large group of known protein

---

⁴For lovers of scientific history, the 1998 biography of this renowned Indian molecular biophysicist (1922–2001) is recommended [1091]. His peppered poetry is an added bonus. (My favorite poem is number 9, on Superhelical Twisting and Replication of DNA [1091, p. 159].)
structures. This superimposed view, averaged over many residues, approximates protein conformational tendencies. The favorable regions correspond to common secondary-structure elements, such as helices and sheets, with finer motifs also noted.

In addition to favorable combinations of $\phi$ and $\psi$ in polypeptides, the side-chain dihedral angle $\chi_1$ has been found to cluster around one of three conformers
known as \( \text{gauche}^+ \) (or \( g^+ \), \( \chi_1 = +60^\circ \)), \( \text{gauche}^- \) (or \( g^- \), \( \chi_1 = -60^\circ \)), and \( \text{trans} \) (or \( t \), \( \chi_1 = 180^\circ \)). These are the favored orientations about tetrahedral atoms (Figure 3.13). Some dependence of \( \chi_1 \) on the residue’s \( \phi \) and \( \psi \) values has also been noted.

### 3.4.4 Conformational Hierarchy

Most natural proteins adopt specific 3D structures that are associated with their biological activity. Of course, proteins are dynamic, but typical thermal fluctuations and local configurational arrangements revolve around a specific globally-folded structure. The majority of proteins is believed to be unknotted in a topological sense, though the polypeptide chain is frequently covalently bonded via disulphide links and noncovalently held together by hydrogen bonds [1254].

One of the hallmarks of biomolecular structure is that the amino acid sequence determines the 3D structure of a protein. This was first shown by Christian B. Anfinsen and his colleagues in the early 1960s [51]. Anfinsen shared the Nobel Prize in Chemistry in 1972 for his work on ribonuclease — connecting the amino acid sequence to the biologically active conformation — with Stanford Moore and William H. Stein — who connected ribonuclease’s chemical structure to its catalytic activity.\(^5\) In Anfinsen’s work, the protein ribonuclease was denatured by destroying its hydrogen bonding network as well as intrinsic disulfide bonds. The researchers observed that the protein spontaneously refolded into its native state in a short time, regaining all its enzymatic activity. Of course, we recognize now that accessory chaperone molecules may be necessary to assist in the folding of many large proteins \( \text{in vivo} \), as discussed in Chapter 2.

Four basic levels are used to describe protein structure:

- **primary structure** — the sequence of amino acids;
- **secondary structure** — regular local structural patterns such as \( \alpha \)-helices and \( \beta \)-sheets, or combination motifs thereof (supersecondary structure);
- **tertiary structure** — the 3D arrangement of all atoms in the polypeptide chain in space; and
- **quaternary structure** (used for large proteins with independent subunits) — the complete 3D interaction network among the different subunits.

The next chapter describes in turn the secondary and supersecondary, tertiary, and quaternary structure of proteins.

---

\(^5\) Readers are invited to browse the electronic museum of Profiles in Science at [www.profiles.nlm.nih.gov/](http://www.profiles.nlm.nih.gov/) for a glimpse not only of Anfinsen’s scientific activities but also of his other hobbies and interests.
3. Protein Structure Introduction
4

Protein Structure Hierarchy

Try to learn something about everything and everything about something.

Thomas Henry Huxley (1825–1895).
4.1 Structure Hierarchy

The complexity of protein structures requires a description of their structural components. This chapter describes the elements of protein secondary structure — regular local structural patterns — such as helices, sheets, turns, and loops. Helices and sheets tend to fall into specific regions in the \{\phi, \psi\} space of the Ramachandran plot (see Figures 3.18 and 3.19). The corresponding width and shape of each region reflects the spread of that motif as found in proteins.

Following this description of each secondary structural element, we discuss the basic four classes of protein supersecondary or tertiary structure (the 3D spatial architecture of a protein), namely \(\alpha\)-proteins, \(\beta\)-proteins, \(\alpha/\beta\)-proteins, and \(\alpha + \beta\)-proteins. This is followed by a presentation of the fold motifs for each such class. Classes and folds are at the top of protein structure classification, as introduced in the last section. Describing these folds and structural motifs is far from an exact science, so variations in some of these aspects are common.

4.2 Helices: A Common Secondary Structural Element

4.2.1 Classic \(\alpha\)-Helix

In the classic, right-handed \(\alpha\)-helix \((\alpha_R)\), a hydrogen bonding network connects each backbone carbonyl (C=O) oxygen of residue \(i\) to the backbone hydrogen of the NH group of residue \(i + 4\) (see Figure 4.1). This hydrogen bonding provides substantial stabilization energy.

The regular spiral network of the \(\alpha\)-helix is ubiquitous in proteins. It is associated with a \{\(\phi\), \(\psi\)\} pair of about \{-60\(^\circ\), -50\(^\circ\}\}. The resulting helix has 3.6 residues per turn, and each residue occupies approximately 1.5 \(\AA\) in length. The helix may be curved or kinked depending on the amino acid sequence, as well as on solvation and overall packing effects. Such distortions are reflected by the \{\(\phi\), \(\psi\)\} distribution around the \(\alpha_R\) region in typical Ramachandran plots.

Hemoglobin, myoglobin, bacteriorhodopsin, human lysozyme, T4 lysozyme, Trp repressor, and repressor-of-primer (Rop) are all examples of proteins that are virtually entirely \(\alpha\)-helical. See Figures 4.2 and 4.3 for illustrations of such \(\alpha\)-proteins (see below) and Figure 3.10 for Rop.

An \(\alpha\)-helix is associated with a dipole moment: the amino terminus of the helix has a positive charge and the carboxyl end has a negative charge clustered about it. Thus, residues that are negatively charged on the amino end and positively-charged on the carboxyl end stabilize the helix; residues with the opposite charge allocation destabilize the helix.

Experimental and theoretical work has shown that both intrinsic and extrinsic (inter-residue interactions) factors are important for helix formation in proteins. Residues with restricted sidechain conformations, due to long or bulky groups, are poorer \(\alpha\)-helix participants than other residues. Glutamine, methionine,
4.2. Helices: A Common Secondary Structural Element

and leucine favor α-helix formation, while valine, serine, aspartic acid, and asparagine tend to destabilize α-helices (e.g., due to steric and electrostatic considerations).

4.2.2 $3_{10}$ and $\pi$ Helices

There are more common variants of the α-helix motif that are typically not stable in solution but can play a part in overall protein structure. These include the tighter $3_{10}$ and looser $\pi$ helices, with $\{\phi, \psi\}$ angles around $\{-50^\circ, -25^\circ\}$ and $\{-60^\circ, -70^\circ\}$, respectively.

The tighter $3_{10}$ helix of three residues per turn (instead of 3.6 in the classic α-helix) involves hydrogen bonds between residues $i$ and $i+3$ instead of $i$ and $i+4$ as in $\alpha_R$. There are 10 atoms within the hydrogen bond; hence the nomenclature $3_{10}$. The more loosely coiled $\pi$ helix has hydrogen bonds between residues $i$ and $i+5$ of the polypeptide.
Because of their close packing, $3_{10}$ helices generally form for a few residues only, often at the C-terminus end of classic $\alpha$-helices where the helix tends to tighten. Similarly, the $\pi$ helix occurs rarely since the backbone atoms are so loosely packed that they leave a hole.
4.2.3 Left-Handed $\alpha$-Helix

A left-handed $\alpha$-helix is theoretically possible, with $\{\phi, \psi\} = \{+60^\circ, +60^\circ\}$. However, this motif is generally unstable. The chirality preference for $\alpha$-helices follows the chirality of L-amino acids.
4.2.4 Collagen Helix

The triple-stranded collagen helix is often considered a specific secondary element. It is associated with \( \{\phi, \psi\} = \{-60^\circ, +125^\circ\} \). A large body of structural data has suggested that extensive hydration networks in the collagen triple helix (among the protein residues and with water molecules) are responsible for collagen stability and assembly (see [115, 680] and references cited therein). A recent hypothesis — that inductive effects by electron-withdrawing residue moieties might play a key factor in collagen’s stability [562] — remains to be proven.

4.3 \( \beta \)-Sheets: A Common Secondary Structural Element

Another common motif is a \( \beta \)-sheet. These sheet regions form by aggregating amino-acid strands, termed \( \beta \)-strands, via hydrogen bonds. Typical lengths of \( \beta \)-strands are 5–10 residues. The aggregation can occur in a parallel or anti-parallel orientation of the strands, as shown in Figure 4.1, each with a distinct hydrogen bonding pattern. Each such \( \beta \)-strand has two residues per turn and can be considered a special type of helix. The hydrogen bond crosslinking between strands — alternating C=O \( \cdots \) H–N and N–H \( \cdots \) O=C — is such that the sheet has a pleated appearance. Thus, in comparison to \( \alpha \)-helices, \( \beta \)-sheets require connectivity interactions that are much longer in range.

For parallel \( \beta \)-sheets, \( \phi \approx -120^\circ \) and \( \psi \approx +115^\circ \). For anti-parallel \( \beta \)-sheets, \( \phi \approx -140^\circ \) and \( \psi \approx +135^\circ \). As for \( \alpha \)-helices, the ring of proline does not adapt well into \( \beta \)-sheets since it cannot participate in the hydrogen bond network between strands. Valine, isoleucine, and phenylalanine have been found to enhance \( \beta \)-sheet formation.

Often, at the edges of \( \beta \)-sheets, an additional residue that cannot be included in the normal hydrogen bonding pattern produces a \( \beta \)-bulge of the extra residue. Figures 4.4 and 4.5 show the structures of proteins that are mostly \( \beta \)-sheets.

4.4 Turns and Loops

Other common structural motifs in proteins are turns and loops.

**Turns** (also called \( \beta \)-turns or reverse turns) occur in regions of sharp reversal of orientation, such as the junction of two anti-parallel \( \beta \)-strands. Such motifs are classified as turns based on distance criteria (e.g., the C\(^\alpha\) atoms of residues \( i \) and \( i + 3 \) are less than 7 Å distant).

**Loops** occur often in short (five residues or less) regions connecting various motifs. Loop regions that connect two adjacent anti-parallel \( \beta \)-strands are known as hairpin loops. Short hairpin loops are found at protein surfaces.
The majority of turns and loops lies on the protein surface because of solvation considerations. They are important elements that allow, and possibly drive, protein compaction. Most loops interact with solvent and are highly hydrophilic (water soluble). Since protein core regions are more stable than short
connective elements of helices and strands, evolutionary differences among homologous sequences are often localized to loop and turn regions. Non-coding regions (introns) are similarly found in genes that correspond to loops and turns in protein structures.
4.5 Formation of Supersecondary and Tertiary Structure

4.5.1 Complex 3D Networks

The secondary structural elements described above often combine into simple motifs that occur frequently in protein structures. Such motifs (or folds) are also called supersecondary structure. Examples are β hairpin (β-loop-β units), Greek key, and β-α-β units (see below).

Supersecondary and tertiary structures of proteins can be described by the specific topological arrangement of the secondary or supersecondary structural motifs. Although the 3D architecture of a protein can be a complex composite of various secondary and supersecondary structural motifs, the majority of the residues — roughly 90% — are found to be involved in secondary structural elements. In fact, on average 30% of the residues are found as helices, 30% as sheets, and 30% as loops and turns. Proteins can be monomeric or multimeric, with subunits that fold in a dependent or independent manner with respect to other domains.

The different polypeptide domains can be connected by disulfide bonds, hydrogen bonds, or the weaker van der Waals interactions. Tertiary structure is also affected by the environment. Hydrogen bonding with solvent water molecules can stabilize the native conformation, and the salt concentration can affect the compact arrangement of the folded chain.

Molecular graphics packages often display the secondary structural motifs clearly by using ribbon diagrams in which helices are depicted as coils and sheets are shown as twisting planes with arrows (see Figures 4.2, 4.3, 4.4, and 4.5, for example).

4.5.2 Classes in Protein Architecture

Based on the known protein structures at atomic resolution, four major classes can be used to describe the arrangement in space of the various secondary structural elements (or domains) of polypeptides:

- **α-proteins** — proteins which form compact aggregates by packing mainly α-helices, often in a symmetric arrangement around a central hydrophobic core;

- **β-proteins** — proteins which pack together mainly β-sheets, with adjacent strands linked by turns and loops and various hydrogen bonding networks formed among the individual strands, often resulting in layered or barrel structures;

- **α/β-proteins** — proteins that are folded with alternating α-helices and β-strands, often forming layered or barrel-like structures; and

- **α + β-proteins** — proteins that combine largely-separated (i.e., non-alternating) helical and strand regions, often by hairpins.
Figures 4.2–4.7 illustrate members of each such class.

Recent statistics for PDB protein structures reveal that approximately 24% belong to the all-α class, 15% to all-β, 12% to α/β, and 32% to α+β. The remaining 17% includes multidomain proteins, membrane and cell-surface proteins, and peptides, and small proteins (see Figures 4.8–4.10). For updated statistical information, check scop.mrc-lmb.cam.ac.uk/scop/, click on ‘Statistics here’. (See last section of this chapter for SCOP description).

Other classes are defined for proteins found on membrane and cell surfaces, small and/or irregular proteins with multiple disulfide bridges, proteins with multiple domains or with bound ligands, and more. Included, for example, are small proteins like rubredoxin (PDB entry 1rb9), various zinc-finger and metal-binding proteins like the cysteine-rich domain of protein kinase (PDB entry 1ptq), disulphide-rich proteins like sea anemone toxin k (PDB entry 1roo), and proteinase inhibitor PMP-C (PDB entry 1pmc).

4.5.3 Classes are Further Divided into Folds

The protein classes are further divided into observed folds for protein structures. Folds describe the arrangement of secondary structural elements and/or chain topology. Each protein class has common folds, as described in turn in the next three sections.

4.6 α-Class Folds

In the α-class of proteins (Figures 4.2 and 4.3), bundles, folded leafs, and hairpin arrays are major fold groups.

4.6.1 Bundles

Bundles occur when α-helices pack together to produce a hydrophobic core. Typically, an array of α-helices is roughly aligned around a central axis. The bundle can be right or left-handed depending on the twist that each helix makes with respect to this axis. A coiled coil (two intertwined helices) can be a building block of these bundles. A simple example of a coiled coil is seen in the DNA-binding leucine zipper protein shown in Figure 6.5 of Chapter 6.

Among the α-protein bundles, the four-helix bundle motif (often written as α₄) is common, as in myohemerythrin, Figure 4.2, and Rop (a small RNA-binding protein involved in replication), Figure 3.10. Other α₄ proteins are ferritin (a storage molecule for iron in eukaryotes), cytochrome c’ (heme-containing electron carrier), the coat protein of tobacco mosaic virus, and human growth hormone.
Multi-helical bundles are also observed in α-proteins; 3–6 and 8-helix aggregates are more frequent than others. Figure 4.2 shows a 5-helix bundle for the transport protein pix.

### 4.6.2 Folded Leafs

Complex packing patterns involving layered arrangements are often features of long α-proteins. For example, in folded leaf folds, a layer of α-helices wraps around a central hydrophobic core. Like bundles, such multihelical assemblies (usually 3 or more) pack together as well as form layers. The longest helices are usually in the center, and often the arrangement contains internal pseudosymmetry.

The globin fold of myoglobin (Figure 4.3) shows such a compact arrangement of a folded leaf arrangement formed by 8 helices, leaving a pocket for heme binding. Cytochrome C6 in Figure 3.12 also displays a folded leaf.

A more complex layered topology is the two-layered ring structure of one α-helical domain in the N-terminal region of the enzyme muramidase in bacterial cell walls, soluble lytic transglycosylate (Figure 4.3). It is built from 27 α-helices, arranged in a two-layered superhelix, leaving a large central hole, thought to be important in its catalytic activity.

### 4.6.3 Hairpin Arrays

Other α-helix assemblies that cannot be described by bundle or folded leaf motifs are often described as hairpin arrays (arrays of α-helix /loop /α-helix motifs). The calcium binding protein calmodulin, for example, has a helix/loop/helix motif where the loop region between two helices binds calcium (see Figure 3.11). Figure 4.2 also shows cellulase cela, a toroid-like circular array composed from 6 hairpins.

An irregular α-protein from an all-α subdomain of the regulator of G-protein signaling 4, namely guanine nucleotide-binding protein, is also shown in Figure 4.3. This protein’s motif contains a 4-helical bundle with left-handed twist and up-and-down topology.

### 4.7 β-Class Folds

Proteins in the β-class display a flexible and rich array of folds, as seen in Figures 4.4 and 4.5. Various connectivity topologies can exist within networks of parallel, anti-parallel, or mixed β-sheets that twist, coil and bend in various ways. Indeed, note the much wider regions of the Ramachandran plot associated with β-sheets than with α-helices (Figs. 3.18 and 3.19).
4.7.1 **Anti-Parallel β Domains**

To describe these intriguing folds, it is simpler to begin with folds associated with the large subclass of β-proteins made exclusively of *anti-parallel* β domains. Such proteins tend to form distorted barrel structures. They can be described in terms of building blocks of two-strand, four-strand, eight-strand units, etc., as follows.

**Two-Strand Units**

The basic two-strand unit, the hairpin (denoted $\beta_2$), involves a $\beta$/loop/$\beta$ motif. It has adjacent anti-parallel $\beta$-strands linked head-to-tail by a turn or loop; see the $\beta$-strands connected as $1 \rightarrow 2$ or $4 \rightarrow 5$ for the head-to-tail direction in fibronectin in Figure 4.4.

**Four-Strand Units**

Proceeding to connections of four $\beta$-strands, there are 24 ways to combine two $\beta$-hairpin units to form a 4-stranded anti-parallel $\beta$-sheet unit. The most common topology is the Greek key (or $\beta_4$). The four strands of a Greek key produce a $\beta$-sandwich through the head-to-tail connectivity of $3 \rightarrow 4 \rightarrow 5 \rightarrow 6$, as shown in the diagrams for fibronectin and *satellite panicum mosaic virus* in Figure 4.4. The $\beta$-strands in these illustrations are labeled according to their connectivity in the protein.

**Eight-Strand Units**

Correspondingly, there are many more ways to combine a larger number of $\beta$-strands from motifs of smaller systems. The two most common folds for 8 anti-parallel $\beta$-strands are jellyrolls ($\beta_8$) and up-and-down $\beta$-sheet.

- The appetizing jellyroll is illustrated in Figure 4.4 for the $\beta$-sandwich coat protein of *satellite panicum mosaic virus*. It is a network of 8 anti-parallel $\beta$-sheets with the connectivity $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$, where strands are shuffled when viewed in the diagram left to right. Note the Greek key submotif in the $4 \rightarrow 5 \rightarrow 6 \rightarrow 7$ subunit of the jellyroll.

- In the up-and-down $\beta$-sheet, each $\beta$-strand is connected to the next by a short loop. It has the simpler connectivity $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 8$, where strands 1 through 8 are written left to right (no shuffling required). Figure 4.4 shows this fold for fatty acid binding protein ($1 \rightarrow 2 \rightarrow \ldots \rightarrow 9 \rightarrow 10$).

4.7.2 **Parallel and Antiparallel Combinations**

More generally, $\beta$-protein topologies made of composites of parallel and antiparallel strands usually form layered or barrel structures. The sandwich, barrel, and $\beta$-propeller are three general reference fold groups.
Sandwiches and Barrels

In sandwiches, β-sheets twist and pack with aligned strands, whereas in barrels the sheets twist and coil so that often the first strand is hydrogen bonded to the last strand to produce closed structures. See the sandwich protein fibronectin and barrel in fatty acid binding protein in Figure 4.4. The immunoglobulins in Figure 3.12(d) are also β-sandwiches where seven strands form two sheets.

Propellers

In β-propeller folds, 6 to 8 β-sheets, each with 4 anti-parallel and twisted strands, arrange radially to resemble a propeller. The 7-bladed propeller of galactose oxidase is shown in Figure 4.5.

Other β-Folds

Other β-folds include β-prisms (3 sheets that pack around an approximate 3-fold axis), barrel/sandwich hybrids (2 β-sheets, each shaped as a half barrel and packing like a sandwich), and β-clips (3 two-stranded β-sheets, forming a long hairpin folded upon itself in two locations). Agglutinin in Figure 4.5 shows a β-prism fold.

Recently, β-helix structures have been identified [238]. The polypeptides contain up to 16 helical turns, each of which contains 2 or 3 β-sheet strands. Unlike the β-sandwiches, the β-sheet strands of a β-helix have little or no twist. Most such β-helix folds known to date are right-handed, as seen in pectin lyase A in Figure 4.5. The β-helix motif has been suggested to occur in the infectious scrapie prion protein [1371].

4.8 α/β and α + β-Class Folds

Even more diverse fold patterns are known for the α/β and α + β-classes of proteins depending on the sheet types (parallel, anti-parallel, or mixed network) and the location of the helices (exterior, interior, or on both faces) with respect to the sheet assembly.

We can broadly classify three fold motifs in this class (see Figure 4.6): barrels — closely packed β-strands (usually 8) with α-helices on the exterior, open structures made of twisted β-sheets (parallel or mixed) surrounded by α-helices on both the exterior and interior, and leucine-rich motifs of curved β-sheets with exterior α-helices in leucine-rich regions.

4.8.1 α/β Barrels

A classic example of a barrel core is the barrel structure of triosephosphate isomerase (TIM), an (α/β)₈ topology (see Figure 4.6). The TIM barrel is one of the most common polypeptide-chain folds known today. TIM’s 8 parallel
β-strands coil to form a central core, and its 8 α-helices pack along the exterior. The central barrel ‘mouth’ is the active site of the protein.

4.8.2 Open Twisted α/β Folds

An example from the highly-variable class of open twisted α/β structures is flavodoxin (Figure 4.6). Note that its helices lie on opposite sides of the β-sheet. Typically, the active sites of proteins in this fold class are near the loop regions that connect β-strands to α-helices. Another member of this class is maltate dehydrogenase, characterized by the Rossmann fold (named after its discoverer Michael Rossmann). This (βαβαβαβ) topology has a central, parallel twisted β-sheet surrounded by α-helices and/or loops. It is an important motif in proteins that bind to nucleic acids.

4.8.3 Leucine-Rich α/β Folds

Ribonuclease inhibitor is an example in the leucine-rich class of α/β folds. Its horseshoe structure is formed by homologous repeats of right-handed β-loop-α units (see Figure 4.6). The 17 parallel β-strands lie on the inside of this horseshoe, with the 16 α-helices clustering on the outside. The leucine residues present in all three segments of the repeating unit — the β-strand, the loop, and the α-helix — pack snuggly together to form a hydrophobic core between the β-strand and α-helix regions.

4.8.4 α+β Folds

Yet more complex fold patterns have been observed for the α+β-class of proteins (see Figure 4.7). This diversity reflects the various topologies of the subdomains (or layers) as well as the richness of connectivity patterns among them.

4.8.5 Other Folds

Examples of multi-domain proteins, membrane and cell surface proteins, and small proteins are shown in Figures 4.8, 4.9, and 4.10.

4.9 Number of Folds

It has been postulated that the number of folding motifs is finite and that the entire catalog of folds will eventually be known with the rapidly-increasing number of solved globular proteins [157, 237, 560]. Such postulates come from stereochemical considerations — for example, there is a small number of ways to link compactly α-helices and β-strands — database analyses, and statistical sampling approaches.
4.10. Quaternary Structure

4.9.1 Finite Number?

The exact number of folds has not been determined. Some studies estimate this number to be several thousand [266,780], while others yield only several hundred [1340,1434] (around 10,000 or 3000 total folds in the former group and 850 total folds in the latter works), so a minimal estimate of around 1000 [1259] and the range of 1000–10,000 seem reasonable [168]. Only time will tell how many folds Nature has produced.

Since many computational folding-prediction schemes use known folds for closely-related sequences or closely-related functions of proteins, a finite number of folds suggests that eventually we will be able to describe 3D structures from sequence quite successfully!

Zhang and DeLisi estimated in 1998 [1434], however, that with the technology available at that time, 95% of the folds will only be determined only in 90 years. They argued that, aside from technological improvements, we should carefully select new sequences for structure determination so as to maximize new fold detection and thereby reduce that time substantially. This is important since the annual number of new folds discovered during 1995–2000 has only averaged around 10%, with even less during 2000–2002. Certainly, careful selection of targets is even more critical if the number of folds is actually larger (e.g., of order 10,000) and associated with single sequence families [266]. The structural genomics initiatives (see beginning of Chapter 2) are certainly accelerating the discovery of new folds [48,214]), but the effect of these projects will take time to assess (see, for example, differing opinions in [87,993]). For updated fold information, search PDB holdings.

4.10 Quaternary Structure

Quaternary structures describe complex interactions for multiple polypeptide chains, each independently folded, with possibly other molecules (nucleic acids, lipids, ions, etc.). The interactions are stabilized by hydrogen bonds, salt bridges, and various other complex intermolecular and intramolecular associations in space. The classic example for a quaternary structure is that of the protein hemoglobin, which consists of four polypeptide chains. The four subunits, each of which contains an oxygen-binding heme group, are arranged symmetrically. Other examples of quaternary structure are DNA polymerases (with catalytic and regulatory components) and ion channels, and protein/nucleic acid complexes with complex structures involving many subunits like viruses, nucleosomes, and microtubules.

4.10.1 Viruses

Virus coats are often comprised of many protein molecules and have intriguing quaternary structures. These protein coats envelope the inner domain which
Figure 4.6. Examples of $\alpha/\beta$-proteins. **TIM** (triosephosphate isomerase) displays an architecture of 8 twisted parallel $\beta$-strands which form a barrel surrounded by $\alpha$-helices. **Flavodoxin**, an electron transport protein that binds to a flavin mononucleotide prosthetic group, displays an open twisted $\alpha/\beta$ fold made of three layers (2 helices at left, 5 $\beta$-strands in the middle, and 2 helices at right). **Maltate dehydrogenase** contains the $(\beta\alpha\beta\alpha\beta)_2$ Rossmann fold in the subunit shown. **Ribonuclease inhibitor**, in the leucine-rich class of $\alpha/\beta$ folds, displays a horseshoe structure.

consists of infectious nucleic acids. For example, the **poliovirus** — a spherical complex of 310 Å in diameter — has a shell of 60 copies of each of four proteins. The coat of **tobacco mosaic virus** combines 2130 identical protein units, each of 158 residues, arranged in a helix around a coiled RNA structure of 6400 nucleotides. This results in a rod-shaped complex 3000 Å long and 18 Å in diameter.
Figure 4.7. Examples of $\alpha + \beta$-proteins: lysozyme, phosphocarrier protein, DNA topoisomerase I, and glycine amidinotransferase.

Figure 4.11 illustrates the structure of the 180-chain tomato bushy stunt virus that infects many plants, including tomatoes and cherry trees. Interestingly, virus coats are assemblies of similar proteins rather than one huge protein or combinations of different proteins, because the relatively small amount of viral nucleic acids must encode this protein coat; at the same time, the nucleic acids must be covered completely. Hence a large protein shell consisting of repetitive motifs satisfies both of these criteria.
Figure 4.8. Examples of multidomain proteins: CbiD-like protein with two domains; FlhC-like protein with three domains; HydB/Nqo4-like with four domains; and Hexameric HIV-1 CA with two domains.

Among the larger molecular structures determined by X-ray crystallography at moderate resolution (i.e., approaching 3.5 Å) is the core particle of bluetongue virus, an agent of disease in both plants and mammals. Its transcriptionally active compartment measures 700 Å in diameter and is composed of two principal structural proteins that assemble in two layers, a core and a subcore, together encapsulating the RNA genome (10 segments of doubled-stranded RNA, ~19,000 base pairs total). The crystal structure revealed how these approximately 1000 protein components self-assemble through a complex mixture of packing mechanisms involved in each of the two layers, using triangulation and geometrical quasi-equivalence packing motifs [484].
4.10. Quaternary Structure

Figure 4.9. Examples of membrane and cell surface proteins and peptides: ATP synthase B chain-like protein, with a long helix; MgtE membrane domain-like protein, with five transmembrane helices; Photosystem I reaction center subunit protein, with three transmembrane helices; and V-type ATP synthase subunit C protein, with nine transmembrane helices.

4.10.2 From Ribosomes to Dynamic Networks

Other examples of quaternary structure are noted for the ribosome, muscle-fiber complexes, bacterial flagellar filaments, and photosynthetic assemblies of membrane proteins.

The *E. Coli* ribosome is a ribonucleoprotein complex with a diameter of about 200 Å constructed from 3 RNA molecules and 55 protein chains [419]. The Nobel Prize in Chemistry was awarded in 2009 to three scientists who independently obtained atomic-level crystallographic views of this magnificent
Figure 4.10. Examples of small proteins: **Coronavirus NSP10-like**, binds two zinc ion per subunit; **Carboxypeptidase inhibitor**, disulfide-rich, $\alpha+\beta$; **Resistin**, disulfide-rich six-stranded $\beta$-sandwich; and **Plant proteinase inhibitor** complexed with calcium and SO$_4$.

RNA/protein machine: Ada Yonath, Venkatraman Ramakrishnan, and Thomas Steitz. For example, the Yonath lab solved the large ribosomal subunit from *Deinococcus radiodurans* [516] and the small ribosomal subunit from *Thermus thermophilus* [1135] (see Fig. 1.1). The Steitz lab reported the structure of the large ribosomal subunit from *Haloarcula marismortui* (2833 of the subunit’s 3045 nucleotides and 27 of its 31 proteins) [85], and Ramakrishnan’s group reported the structure of the small subunit of *T. thermophilus* [1379]. These eagerly awaited structures of the bacterial ribosome were aided by cryo-electron microscopy reconstructions — first reported in 1995 for the ribosome from *E. Coli* (see recent
4.10. Quaternary Structure

Figure 4.11. The structure of **tomato bushy stunt virus**, a spherical arrangement of 180 polypeptide chains, each of 387 amino acids, with every 3 chains making up an asymmetric unit (the subunits are colored blue, green, and red).

views in Figure 1.2 (710)) — which helped crystallographers estimate the initial phasing of their X-ray data (see [171] for a perspective). The combined structural characterizations of the ribosome provided clear evidence that the ribosome is a ribozyme — that is, that the ribosome RNA’s component likely catalyzes peptide bond formation (see Chapter 7, RNA sections).

Muscle cells contain parallel myofibrils composed of two kinds of filaments, each with the following proteins: **myosin** (thick filament), and **actin**,
tropomyosin, and troponin (thin filament); around these filaments, titin — itself two extremely long proteins — plus nebulin form a flexible mesh. Muscle contraction is produced by the interaction of actin and myosin.

The bacterial flagellar motor of the protein flagellin [1085] represents another challenging motor complex solved recently. Filaments of flagellin are formed by an arrangement of stacked flagellin proteins (‘protofilaments’) lined up side by side; an arrangement like loosely rolled sheets of paper results. The remarkable cooperativity among the different filaments leads to conversions between a macroscopic left-handed form — used for swimming — and a right-handed form — used for reorientation of motion. The high-resolution flagellin crystal suggests how this possible structural switch (between left and right-handed supercoiled forms) might occur to direct function.

Insights into the solar energy converters in the membranes of bacteria and plants were provided by the crystal structure of photosystem I, a large photosynthetic assembly of membrane proteins and other cofactors from the thermophilic cyanobacterium S. elogatus [616]. The detailed atomic picture (at 2.5 Å resolution) of the network of 12 proteins subunits and 127 cofactors (chlorophylls, lipids, ions, waters, others) shows the beautiful coordination of all components for efficient absorption and conversion of solar energy into chemical energy.

4.11 Protein Structure Classification

Many groups worldwide are working on classifying known protein structures; see [47, 48, 952, 1259] for a perspective of protein structure and function evolution. Several classification schemes and associated software products exist. A popular program is SCOP: “Structural Classification of Proteins” [887]. (See scop.mrc-lmb.cam.ac.uk/scop/ or connect to SCOP through links available in many mirror sites such as PDB) [262]. These classifications are currently assigned manually, by visual inspection, but some automated tools are being used for assistance.

Also noteworthy is the PROSITE (www.expasy.ch/prosite/) database of protein families and domains intended to help researchers associate new sequences with known protein families. Other databases of patterns and sequences of protein families are PFAM and PRODOM; see [881] for a comprehensive list.

The SCOP levels (top-to-bottom) are: class, fold, superfamily, family, and domain. The sequence, or reference PDB structure, can be considered at the very bottom of this tree.

The top level of the SCOP hierarchy is the class (all-α, all-β, α/β, α + β, multi-domain, membrane and cell-surface, and small proteins). Each class denotes common, global topologies of secondary structure.

Next comes the fold, which clusters proteins that have the same global structure, that is, similar packing and connectivity schemes for the secondary structural elements. Folds are often also called supersecondary structure. From 50 to several
hundred folds are currently known for each class, with the repertoire increasing steadily. An example mentioned above, the α/β barrel fold, groups TIM with other proteins like RuBisCo(C), Trp biosynthesis, and glycosyltransferase into a superfamily, the next level of the classification hierarchy.

The superfamily groups proteins with low sequence identity but likely evolutionary similarity, as judged by similar overall folds and/or related functions. Members of the same superfamily are thus thought to evolve from a common ancestor. Another superfamily, for example, contains actin, the ATPase domain of the heat shock protein, and hexokinase. Superfamilies often pose the greatest challenge in the task of protein classification.

Superfamilies are further divided into families, which cluster proteins with substantial sequence, structure, and function similarity. Generally, this requirement implies a sequence identity of at least 30%, but there are instances of low sequence identity (e.g., 15%) but definitive structural and functional similarities, as in the case of globin proteins. For example, families of glycosyltransferase include β-galactosidases, β-glucanase, α-amylase, and β-amylase.

Finally, at the bottom of the tree of the SCOP classification lies the domain category, to distinguish further structurally-independent regions that may be found in larger proteins.

For updated information on the number of identified folds, superfamilies, and domains, check scop.mrc-lmb.cam.ac.uk/scop/count.html.

As our knowledge of protein structure increases, our classification schemes and software tools will evolve quickly. Automation of the classification is important for rapid structural analysis and ultimately for relating the sequence and structure to biological function.

The reader is encouraged to re-read at this point the sections in Chapter 2 on protein folding/misfolding (Sections 2.2 and 2.3).