CHAPTER 1

Proteins Are Polymers that Fold into Specific Structures

PROTEINS ARE THE MACHINES THAT PERFORM CELLULAR FUNCTIONS

Proteins are biology's workhorse molecules. In a human cell, there are about 18,000 different types of protein molecules. Think of cells as teeming factories of molecules of various types. Proteins perform many functions—as the factory workers, as the machines that produce the factory's output, and as the factory's structural framework. In contrast, nucleic acids (DNA and RNA) are the molecules that encode information, providing the instructions for making the factory products. Sugars and carbohydrates are energy sources used to run the factory. The factory walls (cell membranes) are made up primarily of lipids and polysaccharides. *E. coli* is a bacterium having about 4300 protein-coding genes and a total of about 3×10^6 individual protein molecules [1]. Figure 1.1 shows the size of a typical protein molecule compared with atoms and cells.

Proteins are targets for drug discovery and disease intervention. Discovering new drugs typically involves identifying proteins involved in a disease, and then designing drug molecules or proteins that activate or inhibit those proteins. Some drug molecules are small molecules. Increasingly, proteins are being developed within biotechnology companies, as therapeutic agents to act upon other proteins. In addition, proteins promise to play important roles in nanotechnology, as miniature engines, pumps, motors, optical transducers, and sensors, in

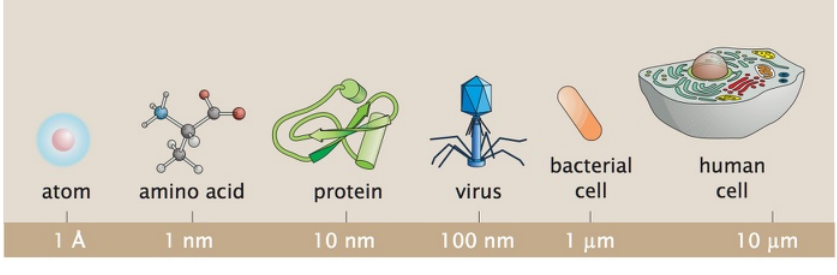

Figure 1.1 In linear dimensions, proteins are a hundredfold larger than atoms and a thousandth the size of a human cell. Note the sizes of images follow the lower size scale and are not drawn strictly to size.

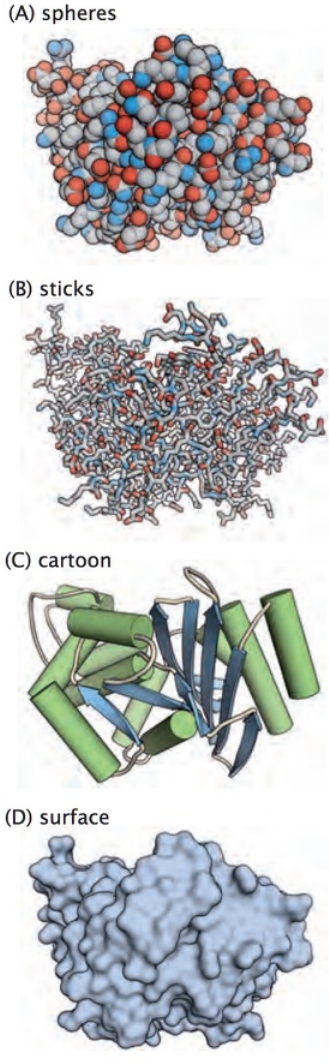

Figure 1.2 Different representations of the same protein structure illustrate different features. (A) A space-filling representation indicates each atom's position, color-coded by atom type: *gray* for carbon, *cyan* for nitrogen, *red* for oxygen, *white* for hydrogen (not shown here), and *yellow* for sulfur. (B) Sticks indicate each bond, colored according to the bonded atoms. (C) Cartoons show secondary structures. (D) Surface renderings show the shapes.

photosynthesis and other applications. Proteins are nature's miniature machines.

To understand the properties of proteins, you first need to understand their structures. Throughout this book, we use different types of images of protein structures (Figure 1.2). Sometimes you want to see chemical and atomic details. In such cases, space-filling models of the atoms or stick representations of the bonds are useful (see Figure 1.2A and B). Other times, you want to see the overall shape of the protein and not the details. In those cases, you may prefer cartoons, or surface representations (see Figure 1.2C and D).

PROTEINS HAVE SEQUENCE–STRUCTURE–FUNCTION RELATIONSHIPS

A protein is a linear polymer molecule. A polymer, like a string of beads or a pearl necklace, is composed of repeat units, called *monomers*,[1] which are covalently linked together in a linear chain-like fashion. In proteins, the repeat units are called *peptides*, so proteins are also called *polypeptides*. There are many types of polymers, natural and synthetic. Synthetic polymers—like commercial polyethylene, polystyrene, or polypropylene—are *homopolymers*: they are repeats of a single type of monomer unit strung together: –AAAAA–, for example. In this respect, proteins are different than almost all synthetic polymers: they are *heteropolymers*. In a protein chain, different types of monomers are strung together in a particular *sequence*: –ILAKW–, for example. For proteins, the units are *amino acids*, also called *residues*, because this is what is left over after the loss of water when two amino acids link together in a chain. There are 20 different types of naturally occurring amino acids. (There are additional types of amino acids, which are occasionally found in proteins, but they are rare.) Think of the set of amino acids as a set of pearls of different colors that are strung together to form a colorful pearl necklace. Or think of the amino acids as forming an *alphabet*. Stringing amino acids together in different linear sequences is a way of encoding information, like the way different sequences of letters encode words in sentences to convey meaning. Proteins have names such as lysozyme, ribonuclease, or barnase. Each protein name describes a chain molecule composed of a particular sequence of covalently bonded amino acids.

The importance of the amino acid sequence of a protein is that it encodes a particular three-dimensional (3D) structure—the protein's *native* or *folded* configuration—into which the protein balls up. In a stunning achievement, the field called *structural biology* has given us atomically detailed structures of more than 100,000 proteins, providing the relative spatial positions of their atoms. Starting with J Kendrew's study of myoglobin in 1958, protein structures have been determined by X-ray crystallography and NMR spectroscopy, and more recently by cryo-electron microscopy. The known structures are collected together in the Protein Data Bank (PDB) [2].

The relationship between protein sequences and their native structures is called the *folding code*. The folding code refers to how the physical

[1] Note that the term *monomer* is sometimes used instead to refer to the individual chains of proteins that have multiple chains. We will use the term *monomer* in both these ways, distinguishing them by their context.

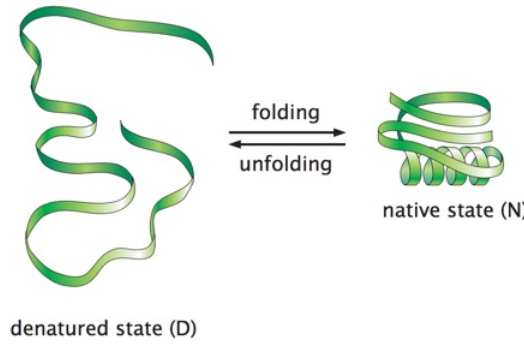

Figure 1.3 The folding process is a transition between denatured (D) and native (N) states, and requires decoding of the information carried in the protein sequence. The denatured state of a given protein is a collection of many different conformations, called an ensemble. Upon folding, each protein collapses into a relatively singular native, or folded, state regardless of its original conformation in the denatured state. The native structure is characteristic of the particular protein. It is significantly more compact than the conformations visited in the denatured state. Proteins fold or unfold depending on environmental conditions.

forces cause a protein's chain of amino acids to fold into only a single stable native structure. A polymer's ability to function because of its folding is mostly limited to proteins and some RNA molecules. A polymer's ability to function because of its folding is mostly limited to proteins and some RNA molecules. No small molecules can encode information in this way. And essentially no synthetic polymers are informational. Synthetic polymers are commercially useful for their material properties—mechanical strength, optical behavior, etc.; they are fibers, fabrics, glues, paints, tires, plastics, films. But, they do not fold into unique structures, and they do not perform the sophisticated types of functions that proteins perform in living systems. And, while DNA molecules can form some structure on the large scales of chromosomes, they have a double-helical structure that is relatively independent of sequence, so they are better suited for carrying information than performing functions. The importance of a uniquely folded structure is that it forces particular chemical groups to sit next to other chemical groups spatially in ways that are stable and functional. These spatial relationships are key to biological mechanisms.

Proteins can undergo *folding* and *unfolding* processes, from the *denatured* or *unfolded* state to the native or folded structure, and vice versa. Think of the unfolded state as a set or *ensemble* of chains that assume a huge number of different 3D arrangements, like a string that can adopt a huge number of different stringy shapes. Upon folding, all chains assume the same conformation, the native structure, but even these can change in limited ways to carry out their functions. The folding process is shown in Figure 1.3 and is described in more detail in Chapters 3, 4, and 6.

The central paradigm of protein science has long been that the amino acid sequence dictates the 3D structure, which, in turn, dictates how the protein performs its function:

$$\text{SEQUENCE} \rightarrow \text{STRUCTURE} \rightarrow \text{FUNCTION}$$

But, this is not the whole story. First, a protein's function can result not just from one structure, but often from transitions between structures.

Protein functionalities can result from large or small *conformational changes*. Second, "structure" does not always mean a single conformation. For one thing, because of the Brownian motion of the solvent, the native protein *wiggles* around its native structure. For another thing, parts of native proteins are often *intrinsically disordered*, whereby the chain is floppy, and does not have a single structure. Third, the biological mechanism of action of a protein can depend not only on its structure, but also on its *dynamics*. So, a more accurate statement of the protein paradigm is

SEQUENCE → STRUCTURES → MOTIONS → FUNCTION

In this chapter, we describe protein structures. In Chapter, 2, we describe protein mechanisms of action. The relationships between motions and function are described in Chapters 10 and 12. We begin with the structures and properties of a protein's building blocks, its amino acids.

AMINO ACIDS ARE THE REPEAT UNITS OF PROTEINS

Figure 1.4 shows the chemical structure of an amino acid, which is the building block unit of proteins. Every amino acid is so-called because it has an *amino* group ($-NH_2$) at one end and a *carboxylic acid* group

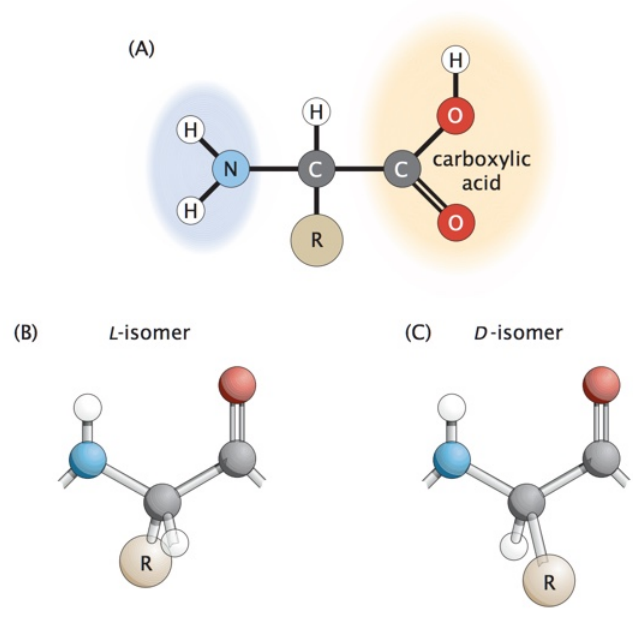

Figure 1.4 The backbone structure of amino acids determines chirality. (A) In an amino acid, the amino group ($-NH_2$ on the left) is connected to the carboxylic acid group (—COOH on the right) through the C^{α}-atom, which is also bonded to the side chain (abbreviated as "R" to be general) and a hydrogen. (B) and (C) 3D view of the repeat units in proteins. Because each C^{α} is tetrahedrally bonded to four different types of atoms, amino acids are either D-isomers or L-isomers. The L-isomer is the naturally occurring form in proteins. The positions of the side chain R differ in the two isomers. In the L-isomer, it points back into the plane of the page, provided that the two flanking backbone bonds $N-C^{\alpha}$ and $C^{\alpha}-C$ lie in the plane of the page and the N-terminal end is on the left (as shown). In the D-isomer, the same arrangement of the backbone leads to an out-of-plane orientation of the R group toward the reader.

(—COOH) at the other. At physiological pH, the amino and carboxylic acid groups are both completely ionized, meaning that one end of the amino acid is a base and the other end is an acid, having a positive and negative charge, respectively. Between the amino and carboxyl ends is a C^{α} atom, also called the α-carbon. Taken together, the amino group, the α-carbon, and the carboxyl group of amino acids are collectively called the *backbone* or the *main chain* of the protein.

Figure 1.4 also shows the *side chain* of an amino acid. The side chain is labeled "R" in the figure as a placeholder to indicate that this can be any one of 20 different types of chemical groups shown in Figure 1.5. One amino acid differs from another in the side chains. The side chain is covalently bonded to the C^{α} atom. The 20 amino acid side chains differ in their chemical structures and physical properties, as described below.

Amino Acids Are Chiral Molecules

Figure 1.4B and C show a property of amino acids called *chirality* or *handedness*. An object is chiral if it does not superimpose on its mirror image. A left hand is chiral because no motion or rotation causes it to look like a right hand. Amino acids are chiral because the α-carbon is tetrahedrally bonded to four different types of groups: carbonyl C, amino N, side chain R, and hydrogen atom H. Glycine is the only natural amino acid that has no chirality, because its R group is just a hydrogen atom, with two of these protons attached to the α-carbon, and for chirality specification these are indistinguishable. There are two different stereochemical ways that the hydrogen and the R group can be situated with respect to the main chain (see Figure 1.4B and C). The two different stereochemical structures are labeled the L-isomer, also known as the left-handed (*levo*) form, and the D-isomer, which is called the right-handed (*dextro*) form. These two different structures are called *optical isomers* because they rotate the plane of plane-polarized light in opposite directions in optical spectroscopy. Biological systems contain L-amino acids almost exclusively. The evolutionary "reason" for biology's use of L rather than D is not known; it may just be a frozen historical accident. The two isomeric forms are chemically quite distinct since the only way to convert between the D- and L-isomers is to break a chemical bond. Later, we discuss how the chirality in the individual amino acids leads to handedness on a larger scale in protein structures.

The 20 Amino Acids Have Different Physical Properties

The 20 different amino acid side chains are shown in Figure 1.5. Each has a different chemical character: *charged*, *polar* (P), or *nonpolar* (also called *hydrophobic*, H). Charged and polar groups have an affinity for water (they are *hydrophilic*), while nonpolar side chains do not. At neutral pH (typical biological conditions), lysine, arginine, and histidine are *bases*; their side chains are positively charged. In contrast, aspartic and glutamic acids are *acids*; their side chains are negatively charged. Alanine, valine, leucine, isoleucine, and phenylalanine side chains are *nonpolar*; they are composed only of hydrocarbon groups (—CH—, $-CH_2-$, $-CH_3$, or aromatics, for example). Nonpolar groups are so-called because even when they are subjected to an applied electric field, they don't become very polarized (that is, they don't develop

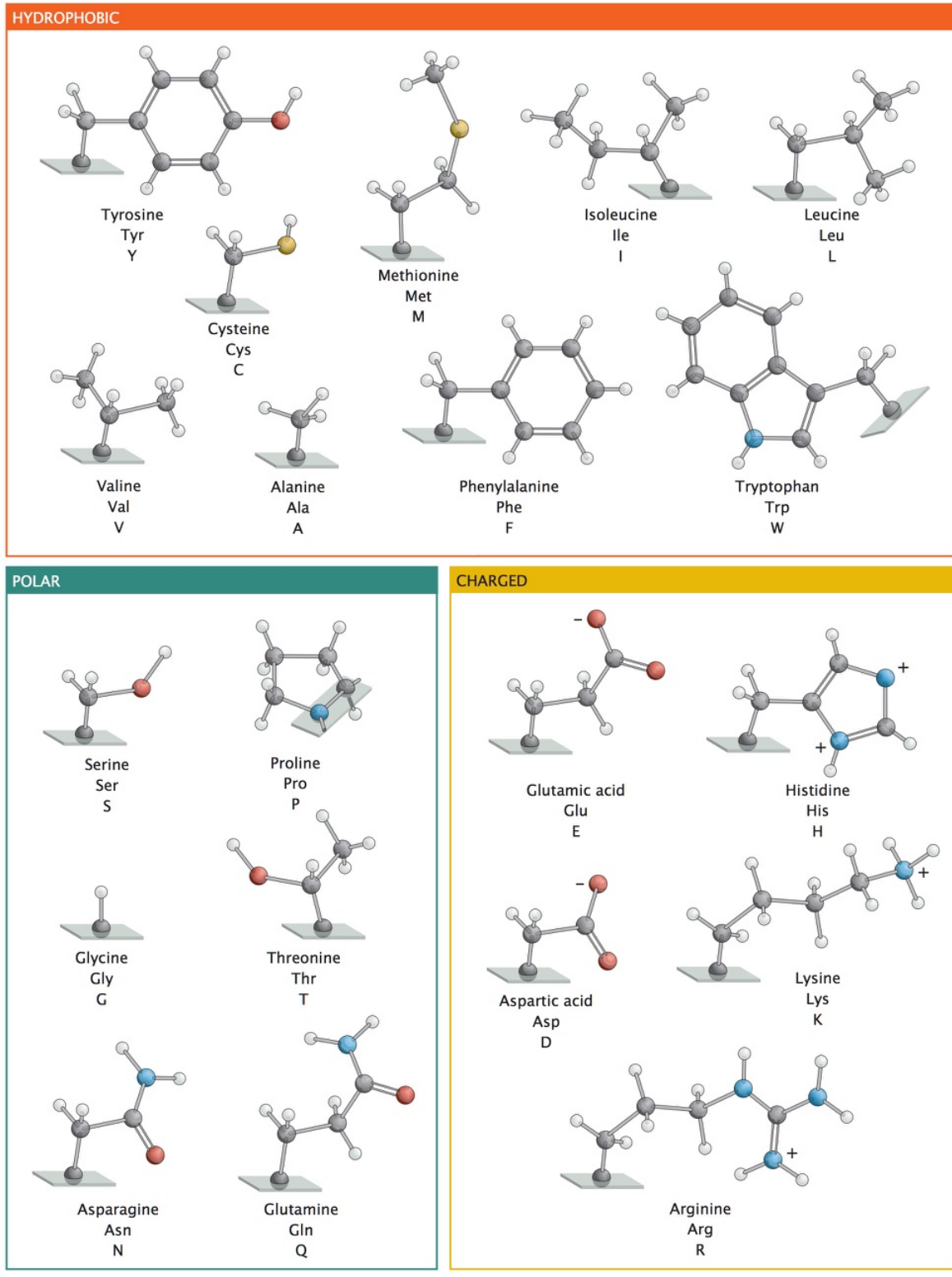

Figure 1.5 Structures of the side chains for the 20 standard amino acids, along with their three-letter and one-letter code names. α-carbons are set in the drawn planes, with the side chains reaching up from them. Amino acid side chains are classified into three broad groups based on their chemical nature: hydrophobic, polar, or charged. The charged residues become uncharged under some conditions of solution pH. Such classifications are from hydrophobicity scales, but there are many of these and some amino acids appear in different categories in different scales.

much internal charge separation or induced dipoles). Nonpolar groups are oil-like; they have an aversion to water. The side chain of glycine is just a hydrogen atom, so the polar groups in glycine's backbone dominate, giving it some limited polar character. Cysteine and methionine, despite their polar sulfur atoms, tend to be buried in the interiors of proteins with the hydrophobic residues, making them appear to be hydrophobic in character. Serine and threonine are polar due to their —OH (hydroxyl) side-chain terminal groups. Asparagine and glutamine are polar due to their amide (—(C═O)—NH_2) side-chain groups. Proline has an imino group fused to the protein backbone and histidine has an imidazole ring.

These physical classifications are neither precisely defined nor always accurate. Some amino acids have multiple personalities, with both nonpolar and polar/ionic characters, depending on their intrinsic pK_a values and the pH of the surroundings. Tyrosine has a bulky hydrophobic part, its phenyl group, but this is attached to a terminal polar —OH. Tryptophan is hydrophobic, but the N—H group on its two-ring indole group can be a hydrogen bond donor. The histidine side chain, a heterocyclic imidazole, can ionize at physiological pH, so histidine can be positively charged or uncharged depending on the pH. Lysine is charged, but it also has a chain of four methylene groups. The broad diversity of structures and biochemical mechanisms of proteins arises from the diversity of the amino acid side chains.

In Proteins, Amino Acids Are Linked Together through *Peptide Bonds*

In a protein, each amino acid is covalently linked to its neighboring amino acid along the chain by a *peptide bond* (Figure 1.6). A peptide bond is a C—N linkage between the carbonyl end of one amino acid and the amino group of the next amino acid.[2] The amino end of the whole protein is called the *N-terminus* and the carboxyl end of the protein is called the *C-terminus*. The standard convention is to number amino acids in a protein chain, $1, 2, 3, \ldots, n$, starting from the N-terminus.

Peptide Bonds Are Planar

The peptide bond, C_{i-1}—N_i, connects amino acids $i-1$ and i. This bond, together with the two backbone bonds on each side, C^{α}_{i-1}—C_{i-1} and N_i—C^{α}_i, lie rigidly within a plane, called the *peptide plane*; see Figure 1.6. The local geometry is planar because the peptide bond has double-bond character (due to the delocalization of the electrons on the carbonyl and amide units). This double-bond character results in rigidity; the peptide bond lacks torsional freedom about its own axis. So, it's possible to think of the backbone as a succession of planes that pivot relatively freely around the C^{α} carbons that join one plane to the next. Figure 1.6A shows a two-amino acid segment of a polypeptide backbone, called a *dipeptide*. Figure 1.6B shows how successive amino acids form peptide planes.

[2]The chemical process of forming a peptide bond is a *condensation* reaction, releasing a water. Thus the formation of a polypeptide containing *n residues* releases $n-1$ water molecules as described by the following polycondensation reaction: $n(\mathrm{HNH{-}C^{\alpha}HR{-}COOH}) \rightarrow \mathrm{H{-}[NH{-}C^{\alpha}HR{-}COO]}_n H + (n-1)\mathrm{H_2O}$.

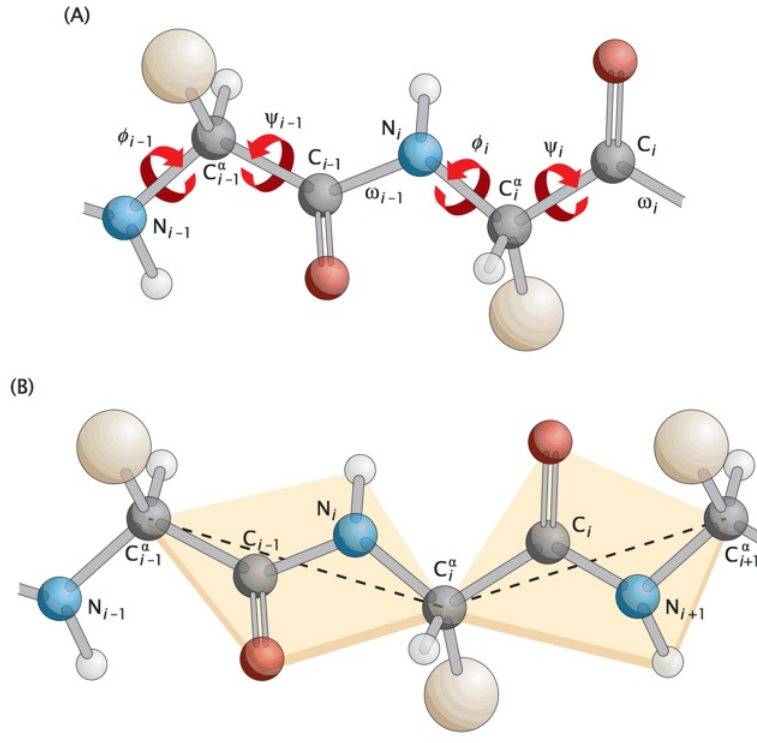

Figure 1.6 Two amino acids in sequence are connected by a peptide bond. (A) Backbone atoms are indexed by residue numbers, $i - 1$ and i. Backbone torsion angles are associated with backbone bonds: ϕ_i with the N_i–C^{α}_i bond, ψ_i with the C^{α}_i–C_i bond, and ω_i with the ith C_i–N_{i+1} peptide bond. Signs of torsion angles correspond to clockwise rotations when looking from the *left* atom of the bond to the *right* atom in (A). (B) Virtual bond model representation of the protein backbone. Their lengths are fixed at 3.8 Å for the usual trans peptide bonds. *Dashed* lines are the virtual bonds connecting successive α-carbons. This representation takes advantage of the rigidity of the peptide bond in the *trans* state and the planarity of the three successive backbone bonds C^{α}_{i-1}–C_{i-1}, C_{i-1}–N_i and N_i–C^{α}_i, along with the corresponding C_{i-1}=O and N_i–H bonds. However, note that the virtual bonds do not have fixed bond angles between them.

One peptide plane can rotate relative to the next peptide plane because of the freedom of the N–C^{α} and C^{α}–C bonds for torsional rotations. Rotation around the N–C^{α} bond defines the torsional (or dihedral) angle called ϕ, while rotation around C^{α}–C defines an angle called ψ. The rotational angle around the peptide bond is called ω (see Figure 1.6A). The ω angle has two stable states called *trans* ($\omega = 180°$) and *cis* ($\omega = 0°$). Figure 1.6 shows the *trans* planar form, which is about 1000-fold more populated than *cis* (except for the bond preceding proline, where the ratio is only about 3-fold). A simplification that is sometimes useful is to represent a chain with *virtual bonds*, shown in Figure 1.6B with dashed lines, which are vectors that join successive α-carbons. These virtual bonds have a fixed length of 3.8 Å for *trans* peptides.

The Rotational Freedom around the Backbone Peptide Bond is Described Using Ramachandran Maps

Different amino acids have different preferred backbone torsional angles ϕ and ψ. These preferences arise from the steric collisions between close neighboring backbone and side-chain atoms [3]. A *Ramachandran map* is a plot of a ϕ angle on the x-axis and a ψ angle on the y-axis for the two bonds that flank a given α-carbon (Figure 1.7). Contours (or colors) on the Ramachandran map indicate the relative populations of the different pairs of angles. Because peptide bonds are usually in their planar *trans* conformation, you can fully specify the locations of the main-chain atoms using only these two dihedral angles, ϕ and ψ, if the bond angles and lengths are known.

Glycine has more freedom than other amino acids in its (ϕ, ψ) angles. Its side chain is only a hydrogen atom, which doesn't collide with the

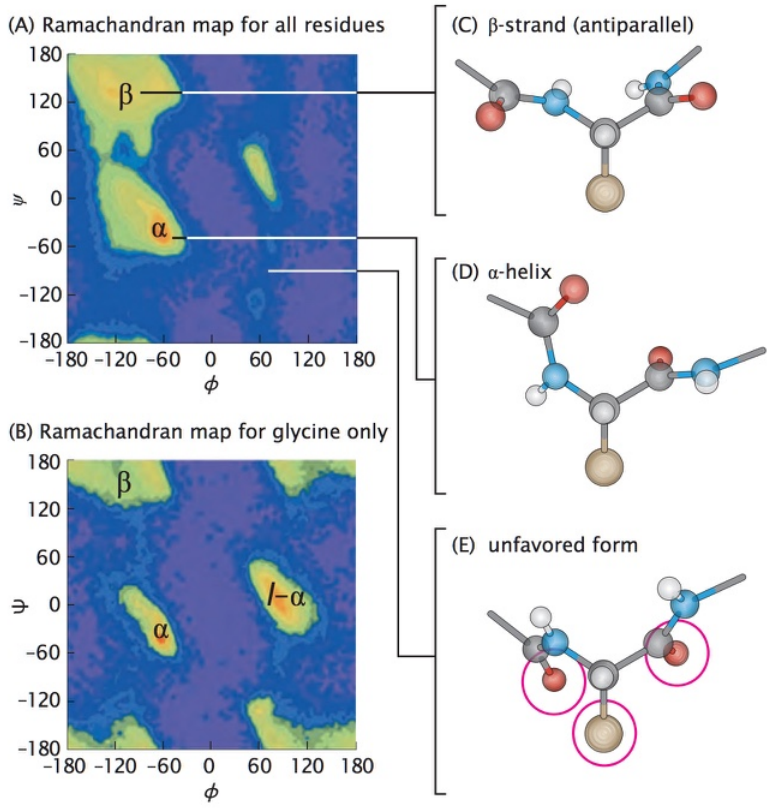

Figure 1.7 A *Ramachandran map* shows the relative populations of different (ϕ, ψ) angle pairs. The data plotted here are from a set of 593 proteins that have a structural resolution of 1.5 Å or better: (A) for all residues and (B) for glycines only. Populations are colored: *red* is the most populated, followed by *orange*, *yellow*, *green*, *blue*, and *violet*. *Violet* regions are disallowed because of steric clashes. Residues other than Gly and Pro tend to fall into one of the three labeled regions: α for right-handed α-helical conformations, β for β-stranded or *l*-α for left-handed α-helical conformations. (C) Illustrates the pair of angles and backbone geometry in the β region; (D) illustrates the local structure in the α-helical region; and (E) shows steric clashes for a left-handed conformation. (The chirality of C^{α} atoms is responsible for the asymmetric distribution of dihedral angles, disfavoring most left-handed ($\phi > 0$) conformations.)

backbone (see Figure 1.7). Proline has less (ϕ, ψ) freedom than other amino acids because its side-chain atoms form a covalently bonded ring with the backbone amide group, locking the backbone ϕ angle.

Side Chains Adopt Preferred Conformations

Side chains, too, have different preferred conformations. Figure 1.8 shows the rotatable bonds on the side chains of tryptophan and lysine. The convention for side chains is to name the heavy atoms by Greek letters as you move away from the C^{α} main chain, that is, C^{β}, C^{γ}, and so on. Similarly, the rotational angles are called χ_1, χ_2, and so on, with subscripts increasing farther away from the main chain.

Why do side chains have favored χ angles? Side chains contain short hydrocarbon chains. So, they have the same conformational preferences that hydrocarbons have. Hydrocarbon chains [for example, $(-CH_2-)_n$] in the gas phase tend to populate three *rotational isomeric states,* called *trans*, *gauche*$^+$, and *gauche*$^-$. Figure 1.9A shows

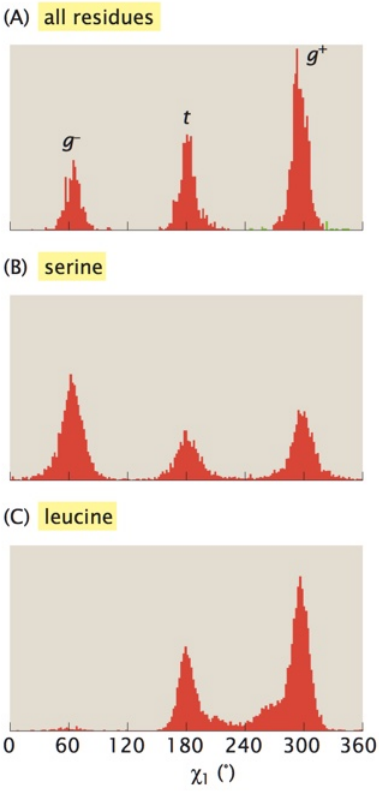

Figure 1.9 Distribution of side-chain χ_1 angles (C^{α}–C^{β} bonds) in proteins and dependence on amino acid type. (A) Frequency of occurrence of χ_1 angles for all residue types. The peaks correspond to the rotameric states, *gauche*$^-$ (g^-), *trans* (t), and *gauche*$^+$ (g^+). The lower probability of the g^- state is due to potential steric clashes between the backbone and side-chain atoms of the bond C^{α}–C^{β}, particularly when the backbone is α-helical. The g^- state is highly probable in serine (B) due to hydrogen bond formation propensity between the side-chain hydroxyl group and the backbone C=O group, but is completely inaccessible in leucine (C) where the branching at the C^{γ} atom gives rise to steric clash with the backbone. (A, from JM Thornton. *Protein Sci*, 10:3–11, 2001 and KS Wilson et al. *J Mol Biol*, 276:417–436, 1998; B and C, from MW MacArthur and JM Thornton. *Acta Crystallogr D*, 55:994–1004, 1999. With permission of the International Union of Crystallography.)

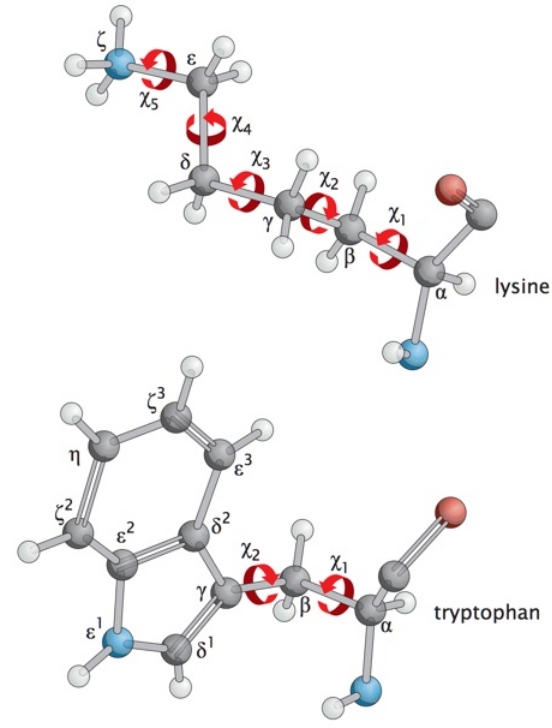

Figure 1.8 Side-chain notation and rotatable bonds of tryptophan and lysine. Side-chain carbon (*gray*) or nitrogen (*blue*) atoms are labeled $\beta, \gamma, \delta, \epsilon, \zeta, \eta$, etc., which indicate the distance (number of bonds) from the main chain C^{α} atom. The torsion angles of the rotatable side-chain bonds are labeled χ_1, χ_2, etc.

that the χ_1 angles in native proteins are the same three conformations. However, Figure 1.9B and C show that the relative populations of the χ_1 angles also depend, to some extent, on the identity of amino acids. While the *gauche*$^+$ state is the most probable rotamer for most amino acids (see Figure 1.9A), serine prefers the *gauche*$^-$ state (see Figure 1.9B). Valine and isoleucine prefer *trans* because of steric restrictions due to branching at their β-carbons. On the other hand, leucine branches at the γ-carbon, leading to steric clashes that disfavor *gauche*$^-$ (see Figure 1.9C).

NATIVE PROTEINS HAVE COMPACT WELL-DEFINED 3D STRUCTURES

Proteins Come in Different Sizes and Shapes

Proteins come in different sizes, as given in terms of either their number of amino acids or their mass. Masses are in units of kilodaltons (kDa). The dalton (Da) is a universal unit of atomic mass. Defined as 1/12 the mass of a carbon atom, it is approximately the mass of a hydrogen atom. The average mass of an amino acid is 136 ± 31 Da. Proteins range in size from *peptides*, which may include up to a few tens of amino acids, to large molecules having thousands of residues. Figure 1.10 shows the distribution of protein chain lengths in yeast, where the average protein size is ~500 residues. In *E. coli*, the average protein has 360 residues. *E. coli's* average protein weighs about 136 × 360 = 50 kDa.

Folded proteins fall into three classes: globular, fibrous, and membrane proteins. Globular proteins are compact and roughly spherical, with axial ratios typically ranging from 1.2 to 1.4 (Figure 1.11). Globular proteins perform cellular functions, including transcription, metabolism, transport, immune responses, cell signaling, and regulation. Fibrous proteins are long and threadlike and serve roles based on their mechanical properties, in structuring cells and tissues, in collagen, silk, hair, and feathers, for example. Globular proteins are usually soluble in water, whereas fibrous proteins are not. Membrane proteins are located in membranes, an oil-like environment. Membrane proteins can transport molecules and allow the flow of ions across membranes, relay signals across membranes, or perform enzyme activities. We first focus on globular proteins.

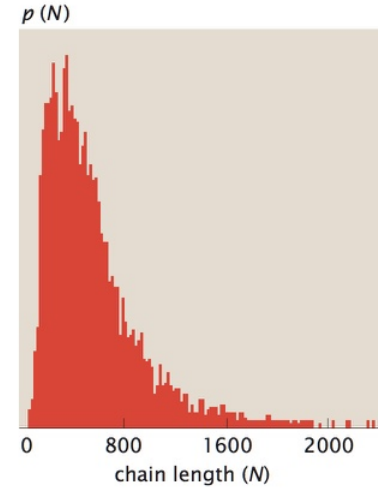

Figure 1.10 Size distribution of proteins expressed in the yeast *Saccharomyces cerevisiae* proteome (set of all proteins). The average length is 501 residues. (From J Warringer and A Blomberg. *BMC Evol Biol*, 6:61, 2006.)

Native Protein Chains Are Balled Up and Tightly Packed

There is not much free space inside a folded globular protein. Proteins are well packed. The density of a protein's atoms in its folded state is about the same as the density of atoms in liquids and solids, or about the same as the density of marbles in a jar. One way to characterize the tightness of packing is to measure the volume occupied by the individual amino acids in folded structures, and compare that volume with the total volume of the protein's constituent atoms. Even liquids

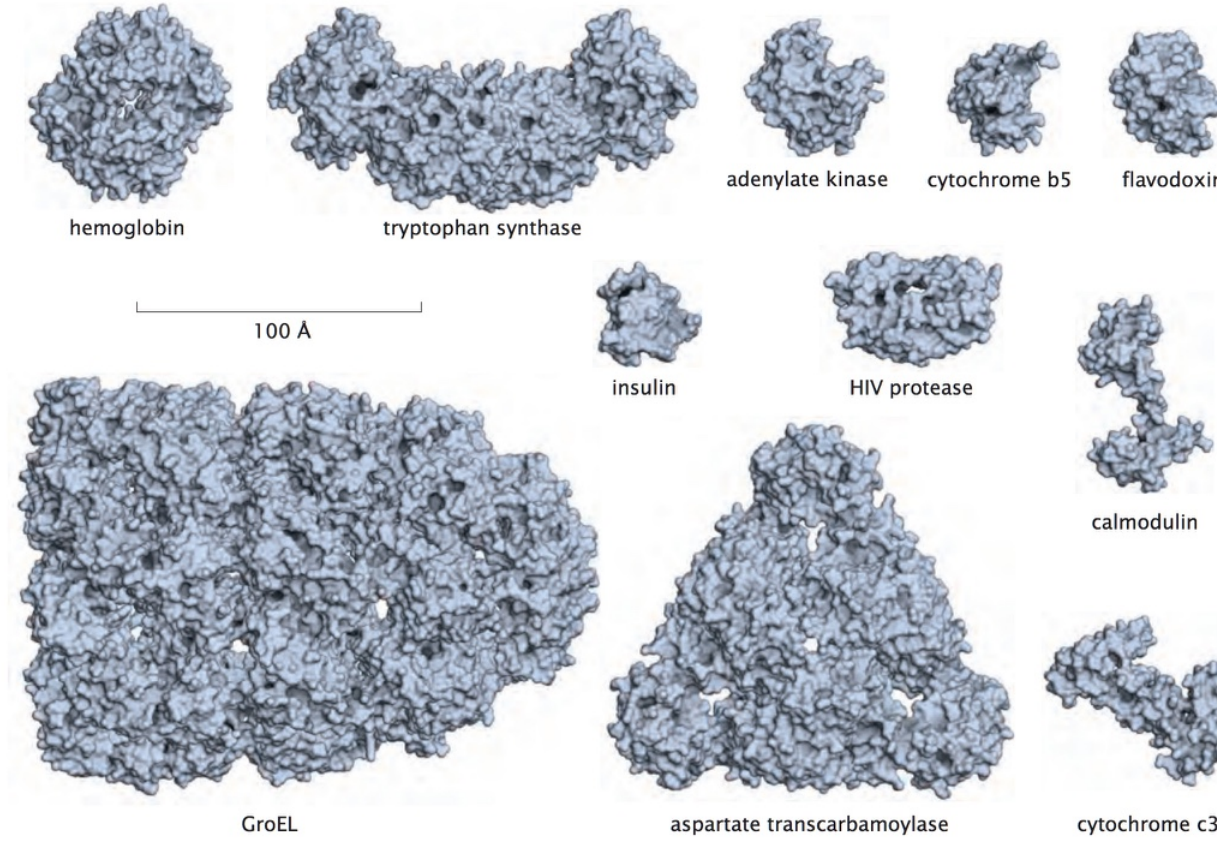

Figure 1.11 Surface representations of eleven proteins illustrate the range of their sizes and shapes. The proteins are all shown at the same magnification.

Table 1.1 Physical properties of amino acids

Residue	van der Waals volume (Å^3)[a]	Average packing density in proteins[b]	Occurrence[c] (%)
Ala	67	0.73	7.5
Arg	148	0.73	5.2
Asn	96	0.77	4.6
Asp	91	0.73	5.2
Cys *(disulfide)*	86	0.81	1.8[d]
Cys *(thiol)*		0.73	
Gln	114	0.71	4.1
Glu	109	0.70	6.3
Gly	48	0.73	7.1
His	118	0.71	2.2
Ile	124	0.73	5.5
Leu	124	0.74	9.1
Lys	135	0.79	5.8
Met	124	0.73	2.8
Phe	135	0.67	3.9
Pro	90	0.70	5.1
Ser	73	0.74	7.4
Thr	93	0.76	6.0
Trp	163	0.68	1.3
Tyr	141	0.69	3.3
Val	105	0.74	6.5

[a] Hard core volume of constituent atoms (data from FM Richards. *J Mol Biol*, 82:1–14, 1974).
[b] Calculated by dividing the van der Waals volume by the average volume of buried residues (that is, those with less than 5% of possible surface area accessible to solvent) for each residue (data from C Chothia. *Nature*, 254:304–308, 1975 and TE Creighton. *Proteins: Structures and Molecular Properties*, 2nd ed. WH Freeman, NY, 1993).
[c] Amino acid frequency from 1021 unrelated proteins (data from P McCaldon and P Argos. *Proteins*, 4:99–122, 1988).
[d] Frequency for all Cys residues.

and solids have some crevices between atoms. In the tightest packing of perfect spheres, only 74% of the volume is filled—the rest is empty cavities. Table 1.1 shows that the packing density of amino acids buried in globular proteins is also 0.74. And, typically, less than about 3% by volume inside a protein is filled by water [4]. However, while a protein core is packed *tightly*; it is not packed *uniformly*, because side chains are different, diverse, and irregular. A protein interior more closely resembles a jar of nuts and bolts than one of marbles.

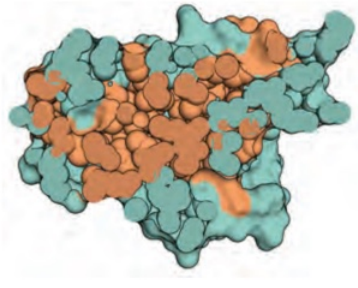

Figure 1.12 Proteins have a hydrophobic core. A cross section of interleukin-4 exposes the protein interior. Hydrophobic Trp, Phe, Tyr, Leu, Ile, Cys, Met, and Val residues are *orange*, and all others are *teal*. Hydrophobic residues tend to be buried, and polar residues lie mostly on the surface.

Proteins Have Hydrophobic Cores

In a folded protein, some amino acids are on the surface and some are buried in the protein's *core*. The surface of a water-soluble protein contains a mix of polar, charged, and nonpolar groups. As shown in Figure 1.12, the core contains mostly nonpolar amino acids, implying that a key force for protein folding is the tendency of the oil-like amino acids to cluster together to avoid contact with water. Oil and water don't mix. So a protein folds up, like an oil droplet, in such a way that its nonpolar amino acids are buried in a core, and the rest of the amino acids are on the surface. We explore the nature of the forces that drive protein folding in Chapter 3, but we introduce the types of interactions

briefly here because they have a prominent role in defining structures and mechanisms.

The Amino Acids in Native Proteins Are Hydrogen-Bonded to Each Other

A folded protein chain has extensive hydrogen bonding. A hydrogen bond is a noncovalent interaction between two chemical groups that can share a hydrogen atom. The two sharing groups are called the hydrogen-bond *donor* and *acceptor*. A folded protein contains extensive networks of C═O ··· H—N backbone hydrogen bonds between different amino acids. The amide group is the hydrogen-bond donor and the carbonyl group is the acceptor. The geometry from donor to hydrogen and from hydrogen to acceptor tends to be nearly collinear.

Proteins can also form hydrogen bonds with water. Water molecules can donate hydrogens or accept hydrogens. Hydrogen bonds can also form between the side chains or between the backbone and the side-chain atoms of proteins. Hydrogen bonds are prominent features in the most common protein substructures, called α-helices and β-sheets.

Cysteines Can Form Disulfide Bonds

Cysteine is an amino acid that terminates in a *sulfhydryl* (also called *thiol*) group (—SH). Under oxidizing conditions, two different cysteine side chains can form a covalent bond with each other, called a *disulfide bond* (—S—S—). (Disulfide bonds can be broken by adding a reducing agent to a protein solution.) Disulfide bonds act as cross-links either within a protein, or between two proteins, and can impart thermal and mechanical stability.

PROTEINS HAVE HIERARCHIES OF STRUCTURE

Proteins have various forms of internal structure. The levels of structure were named primary, secondary, and tertiary by Kaj Linderstrøm-Lang [5] (Figure 1.13). The *primary structure* of a protein refers to its linear sequence of amino acids. The *secondary structure* refers to *helices* and *β-sheets*, the two main types of hydrogen-bonded regular (ordered) substructures in proteins. In an average protein, about 60% of the residues participate in α-helices or β-strands. The next higher level of organization, a protein's *tertiary structure*, refers to the 3D arrangement of its secondary structural elements and its connecting turns, loops, or coiled segments. Protein tertiary structures are described by terms such as "four-helix-bundle" or "β-barrel," reflecting the packing geometry of secondary structural elements. Tertiary structures are stabilized by *tertiary contacts* (generally hydrophobic interactions) between amino acids that are distant in the sequence but close in space.

There are also intermediate levels of structure between secondary and tertiary structures, called *structural motifs*, or *supersecondary structures,* as will be discussed later. *Quaternary structure* refers to the 3D organization of multiple chains, also called *subunits* or *monomers*, for proteins that are composed of more than one chain (here, monomer refers to one polypeptide chain, not to be confused with the amino acid

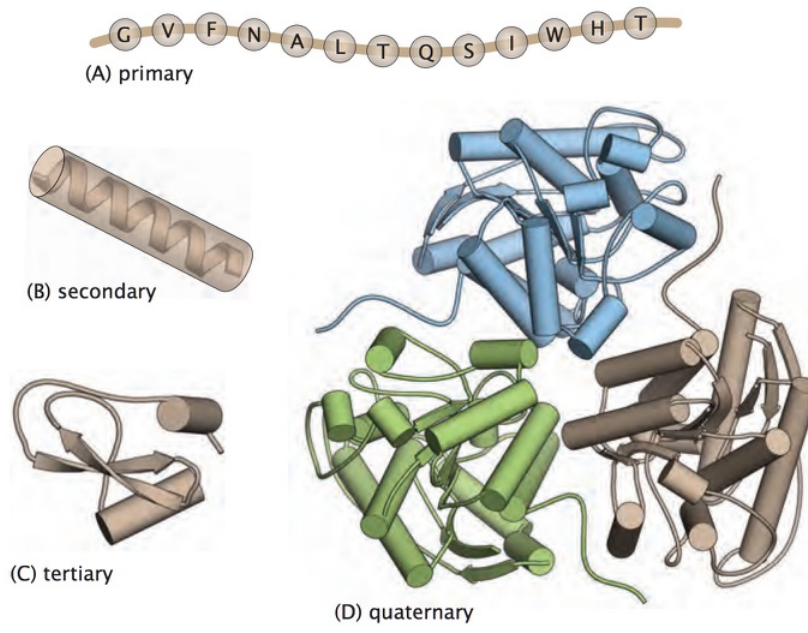

Figure 1.13 There are four hierarchical levels of structure in proteins. (A) A protein's primary structure is its amino acid sequence. (B) The secondary structures are the helices and sheets. Here, the α-helices are represented as cylinders. (C) The tertiary structure is the arrangement of all the structural elements in a single protein chain (beta-strands shown as arrows), while (D) the quaternary structure is the assembly of two or more chains within a protein. Here we show the quaternary structure of arginase (a trimeric enzyme in the urea cycle, which allows the body to dispose of ammonia).

monomers). The quaternary structures of multichain (also called *multimeric*) proteins are usually held together by noncovalent interactions between the individual *subunits*. For example, hemoglobin has a quaternary structure; it is composed of four subunits, or monomers, which are symmetrically assembled (but not covalently bonded) to form a tetrameric structure. Fibrous proteins often have quaternary structures that are stabilized by covalent bridges.

The Secondary Structures of Proteins Are Helices and Sheets

α-helices. A major component of protein structures is the *α-helix*. In an α-helix, the protein backbone spirals around its long axis (Figure 1.14). Each helical turn is composed of 3.6 amino acids, and the helix has a pitch of 5.4 Å between successive turns (or 1.5 Å rise [projected distance along the helix axis] per amino acid). There is very little empty space in the center of the helix. The broad occurrence of the α-helical structure arises from two sources. First, it is an energetically accessible conformation for these (ϕ, ψ) angles (see Figure 1.7). Second, the α-helix is stabilized by hydrogen bonds between the carbonyl oxygen of amino acid i and the amide hydrogen of amino acid $i+4$ (see Figure 1.14). α-helices can be formed by any of the amino acids (except proline) because the hydrogen bond donors and acceptors are backbone atoms.

Interestingly, the existence of the α-helix was hypothesized by Linus Pauling and his colleagues in the early 1950s, before it was discovered in nature [6]. At that time, it was not surprising that a polymer would have a helical structure. Many types of linear polymer chains tend to form various helical structures. Think about a repeated string of vectors with a fixed torsional angle between them. If that angle is 180°, the chain will be a linear rod. But for any other angle, the polymer will be a string of repetitive twist steps. That defines a helix. Typical nonbiological polymers have a single favored repeat angle (at low temperatures), so they crystallize into helical structures. About half of

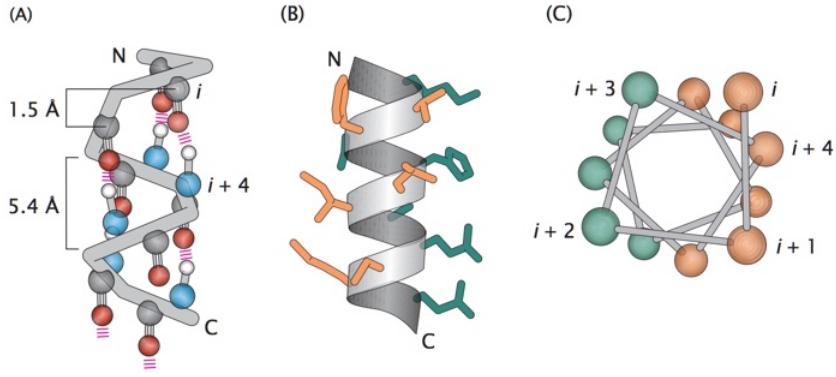

Figure 1.14 Different representations of the α-helix structure in proteins. (A) The regular main chain path is maintained by hydrogen bonds between the carbonyl O of residue *i* and the amide H of residue *i* + 4, represented here with *red* springlike connections. Successive α-carbons occupy angular positions at 100° intervals around the helical axis (to give 3.6 residues per full turn of 360°). Note that the first three N—H groups and the last three C=O groups are not hydrogen-bonded. (B) While the path of the main chain in an α-helix is fixed, the side chains (shown here as sticks) are more free to rotate. Frequently, helices are amphipathic, with hydrophobic residues (*orange*) on one side and hydrophilic residues (*teal*) on the other. (C) The amphipathic character is illustrated in a helical wheel diagram, showing the side chains as spheres viewed around the central helical axis.

the known polymer crystal structures take on one of the 22 different types of helices [7]. The surprise in Pauling's correct prediction was that protein helices would have a noninteger number of monomers per turn, 3.6 in this case. Pauling's key insight was that peptide helices would be stabilized by the hydrogen bonding between the backbone units, from the carbonyl group of one amino acid to the amide group of another.

Helices have a property called *handedness*. A helix can be either right-handed or left-handed, based on which direction it spirals (Box 1.1). Many helices in globular proteins also have a "sidedness" property. Called *amphiphilic* or *amphipathic*, those helices have a stripe of mostly hydrophobic amino acids down one side and a stripe of mostly hydrophilic amino acids down the other side. The hydrophobic stripe usually faces inward in a protein structure toward the protein core. A simple device for visualizing such patterns down the lengths of helices is a *helical wheel* diagram, shown in Figure 1.14C.

Box 1.1 Defining the *Handedness* in a Helix

Here's how a right-handed helix is defined. Align your right thumb so that it points along the helical axis that runs from the N-terminus toward the C-terminus. A helix is right-handed if it curls in the same way as your fingers curl from your palm through your fingers to your fingertips. Otherwise, it is left-handed.

In natural proteins, α-helices are right-handed (Figure 1.15). You can understand this from two basic facts: (1) side chains are on the outsides of helices (see Figure 1.14B) and (2) the naturally occurring amino acids are the L-isomers (see Figure 1.4). The L-amino acids prefer to form a right-handed helix because that minimizes steric

conflicts between the side chains of L-amino acids and the helical backbone carbonyl groups around the outside of the helix (see Figure 1.7E).[3] In synthetic proteins that are made out of D-amino acids, the helices are left-handed.

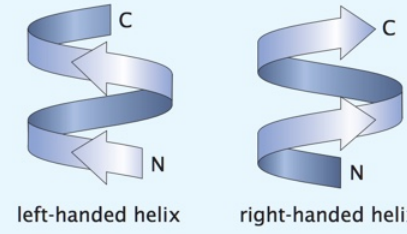

Figure 1.15 The contrast between a right- and a left-handed helix.

β-sheets. In 1951, Linus Pauling also predicted another type of secondary structure before it was found in nature [8]. *β-sheets* are train-track-like structures in which two or more segments, called *β-strands*, are aligned side-by-side (Figure 1.16). Pairs of strands are held together by amide-to-carbonyl main-chain hydrogen bonds from one strand to another. β-sheets are called *parallel* if the N → C directions of both strands are the same, or *antiparallel* if strands run alternately in opposite directions to each other. The side chains lie above or below the plane of the sheet, where they can interact with side chains from adjacent strands as well as those on neighboring sheets above or below. β-sheets are therefore intrinsically stabilized by (i) hydrogen bonds; (ii) side-chain interactions, often of nonpolar groups; (iii) favorable (ϕ, ψ) angles in the β-region of the Ramachandran map (see Figure 1.7); and (iv) by the good packing they achieve. Good packing is stabilized by so-called *van der Waals attractions*.

Less common types of secondary structure. α-helices and β-sheets are the most common—but not the only—types of secondary structure in proteins. Another type is the π-helix, in which a hydrogen bond is formed between residues $(i, i+5)$, instead of $(i, i+4)$ as in the α-helix. Another example is the 3_{10}-helix, in which hydrogen bonding is between residues $(i, i+3)$. The 3_{10}-helix has three residues per turn and ten atoms in the ring closed by the hydrogen bond. It is sometimes found in short peptides and occasionally in proteins, but its hydrogen bonds are less stable, its side-chain packing is less favorable, and its dipoles are more poorly aligned than the hydrogen bonds in the α-helix. The π-helix is rare because its rise is short, bringing side chains into close proximity. Another type of helix, called the polyproline II helix, is found in chains having high proline content, such as collagen (Table 1.2).

[3] This can be rationalized as follows: the carbonyl oxygen is larger than the amide hydrogen. Steric clashes between the side chain and the carbonyl oxygen restrict the range of possible pairs of (ϕ, ψ) values. In particular ϕ values in the range $\phi_i < 0°$ bring the R_i and $(N{-}H)_{i+1}$ groups into close proximity without causing an overlap of their atoms, while values of $\phi_i > 0°$ cause an overlap between R_i and the larger polar group $(C{=}O)_{i-1}$. The lower right quadrant of the Ramachandran map (see Figure 1.7) is almost entirely excluded due to the steric clash between R_i, $(C{=}O)_{i-1}$ and $(N{-}H)_{i+1}$.

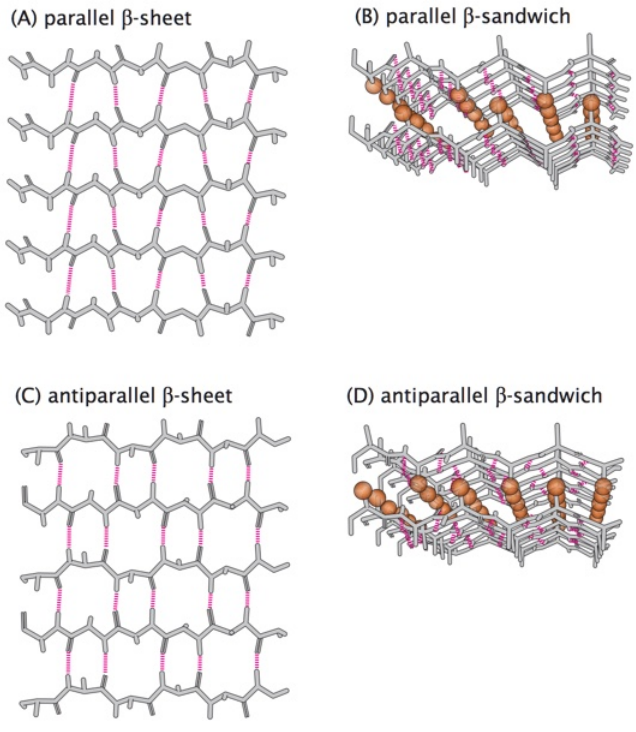

Figure 1.16 The structures of parallel β-sheets and sandwiches. The *left* column shows the hydrogen-bonding patterns in (A) parallel and (C) antiparallel β-sheets. On the *right*, pairs of sheets are tipped down so you can see the hydrophobic side-chain contacts sandwiched between the (B) parallel and (D) antiparallel sheets. (Adapted from JD Schmit, K Ghosh, and KA Dill. *Biophys J*, 100:450–458, 2011.)

Turns and loops. *Turns* and *loops* are not themselves secondary structures.[4] Turns are chain segments of three to five residues usually folded in well-defined geometries and stabilized by local hydrogen bonds. Loops are longer segments of chain that are typically less structured. Turns, also called *reverse turns*, are places where the chain reverses direction. Reverse turns tend to be composed of polar residues and glycine and proline, while loops tend not to have such preferences. Because they are polar, reverse turns tend to be located on protein surfaces. Glycine is common in turns because it contorts easily, due to the lack of side-chain hindrance. Proline is also common in turns because it has the unusual feature that its backbone is constrained in a chemical-ring structure.

Sometimes the N-terminus of a helix has an *end-capping* hydrogen bond. At the ends of helices, the backbone carbonyl and amide groups may have unsatisfied hydrogen bonds. In such cases, a side chain (for example serine) can form a hydrogen bond by folding back onto the backbone to hydrogen-bond with the unfulfilled backbone hydrogen bond.

[4]Sometimes, turns and loops are included in the definition of secondary structures. In this book, we use the term secondary structure only to refer to the regular repeating structures—helices and sheets.

Table 1.2 Geometric parameters for some regular protein conformations

Secondary structure	Residues dihedral angle (°)			# of amino acids per turn	Rise per residue (Å)
	ϕ	ψ	ω		
Right-handed α-helix	−57	−47	180	3.6	1.50
Left-handed α-helix	+57	+47	180	3.6	1.50
3_{10}-helix	−49	−26	180	3.0	2.00
π-helix	−57	−70	180	4.4	1.15
Polyproline II (left-handed)	−82	147	180	3.0	3.04
Parallel β-sheet	−119	113	180		
Antiparallel β-sheet[a]	−139	135	−178		
Fully extended chain[b]	180	180	180	2.0	3.6

Partly adapted from JC Kendrew, W Klyne, S Lifson, et al. *Biochemistry* 9: 3271–3479, *J Biol Chem*, 24:6489–6497, 1969, and *J Mol Biol* 52:1–17, 1969; GE Schulz and RH Schirmer. *Principles of Protein Structure*. Springer-Verlag, New York, 1979.
[a] Twisted β-sheets are abundant in proteins. There are considerable variations among the dihedral angles of twisted β-strands.
[b] Included for reference only. Fully extended chains are not commonly observed and do not form stable sheet-like chain organizations.

Supersecondary Structures, Also Called *Structural Motifs*, Are Common Combinations of Secondary Structures

Some assemblies of secondary structures appear so frequently that they are given names. As a class, they are called *supersecondary structures*. A prominent example is the *coiled coil*, in which two helices are twisted together. Another is the β-helix formed by the association of β-strands in a right-handed or left-handed helical pattern. The β-helix structure is highly stable. Notably, triple-stranded β-helical structures function as cell-puncturing devices of bacteriophages [9]. Their high stability and helical structure are essential to disrupting the host cell membrane during infection.

Supersecondary structures often have a *right-handed twist*. For example the *EF hand*, the β-hairpin, and the β–α–β motif usually exist in the right-handed form [10]. Ninety-five percent of β–α–β motifs in proteins are right-handed [11]. The chirality of supersecondary structures refers to the relative rotational orientation of these structures with respect to the chain axis, as we move along the sequence from the N- to the C-terminus. Figure 1.17 illustrates the differences between a left-handed and right-handed β-sheet and between helical motifs. The prevalence of right-handed twists for individual β-strands is explained by the intrinsic preferences of L-amino acids. Take an individual strand, with right-handed twist, and fold it into a compact structure, a loop, for example. If the loop is right-handed, it will naturally release the right-handed twist/strain of the strands. If the loop were left-handed, it would increase the strain. You can see this by first twisting a belt and then forming it either into a right-handed or left-handed loop. One way releases strain and the other increases it.

Some Substructures of Proteins Are Compact Functional *Domains*

Another type of protein substructure is called a *domain*. A domain is a piece of a polypeptide chain that can fold on its own, function on its own, evolve on its own, or can be identified by its compactness

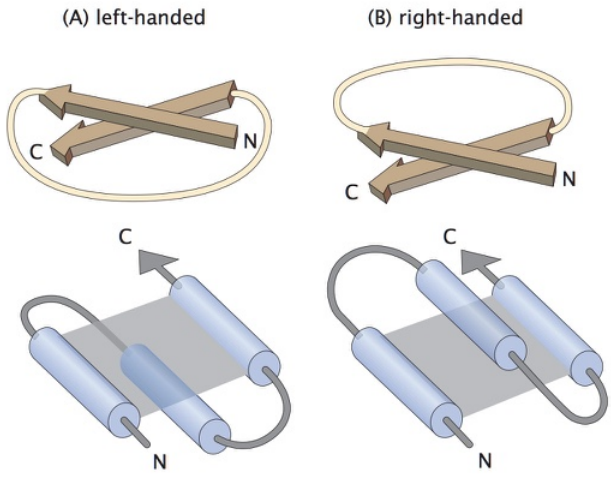

Figure 1.17 Chirality in supersecondary structures. Parallel β-strands or helical motifs may be (A) left- or (B) right-handed. The right-handed path relieves the strain of right-handed β-strands. (Adapted from JM Thornton. *Protein Sci*, 10:3–11, 2001.)

relative to neighboring parts of a chain. A domain is the smallest unit of protein function. Examples include the SH3 and ATPase domains. Often each domain has a *modular function* to perform in a protein, such as binding a ligand, spanning a cell membrane (transmembrane domains of membrane proteins), containing the active site (catalytic domains in enzymes), binding a nucleotide (DNA/RNA-binding domains in transcription factors), or providing a surface for binding other proteins (substrate-binding domains). A single protein chain may have one or more domains, often several. For example Figure 1.18 shows that pyruvate kinase has three domains. Sometimes evolution reuses domains; the same domain can appear in different proteins. Interestingly, the different domains within a protein are sometimes encoded by different regions within the genome, as a result of genetic recombination.

As of 2015, there were known structures of more than 170,000 distinct domains, ranging from 13 to more than 1000 residues.

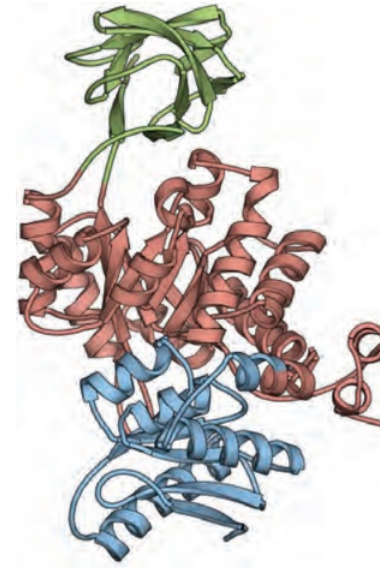

Figure 1.18 Protein domains are independently folding subunits. Pyruvate kinase consists of three domains formed by the N-terminal (*red*), PK domain (*green*), and C-terminal (*blue*) residues.

Native Protein Topologies Are Described Using *Contact Maps*

A useful way to visualize a protein's *topology*—its substructures and their positions relative to each other—is to use a *contact map* (Figure 1.19). For a chain having n residues, the contact map is an $n \times n$ matrix in which the element (i, j) is 1 (or a dot) if residues i and j are in contact, and 0 (or empty) otherwise. Contact maps are symmetrical, so usually only the upper (or lower) diagonal half of the matrix is displayed.

A contact map provides information on all inter-residue contacts in a protein's tertiary structure. In particular, secondary structures appear as simple patterns. For example, helices appear as lines of dots adjacent and parallel to the main diagonal. Parallel β-strands are also parallel, but not adjacent, to the main diagonal on the contact map. Antiparallel strands are lines of dots perpendicular to the main diagonal. Contacts are called *local* if they are near the main diagonal, and *nonlocal* if they are more distant. Helices and turns involve only local contacts, whereas sheet contacts are predominantly nonlocal. Figure 1.19 shows the contact map for the native structure of the chymotrypsin inhibitor (CI2), which has all of these elements.

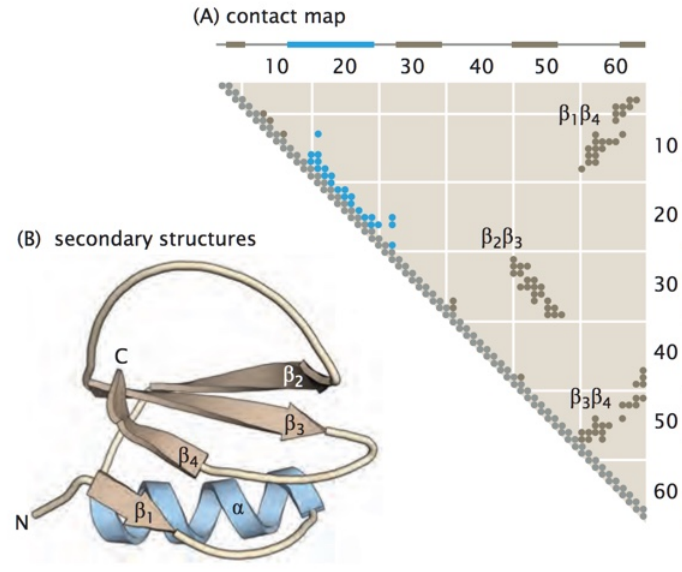

Figure 1.19 (A) Contact map for chymotrypsin inhibitor (CI2). The two axes represent the amino acid number along the sequence. A contact between two residues (represented by a dot on the map) is defined here whenever C^{α} or C^{β} atoms from different residues are within 6 Å of each other. The four large clusters of contacts indicate the main structural interactions in the protein. (B) The 3D structure and identity of β-strands in CI2. The helix is shown in *blue* and strands in *brown* in both (A) and (B). (Adapted from C Merlo, KA Dill, and TR Weikl. *Proc Natl Acad Sci USA*, 102:10171–10175, 2005.)

How Can You Classify Protein Tertiary Structures?

Suppose you discover a new protein structure and want to know if it resembles other known structures. You need a way to classify its structure, starting from its most global features and working down to its most detailed features. Proteins were first classified into four structural families in the 1970s [12]: α (mostly α-helical), β (mostly β-sheet), α/β (sequentially interspersed α-helices and β-strands), and α + β (one region of mainly α-helices joined to another region of mainly β-sheets). Now that many more protein structures are known, these initial four structural classes have become the basis for more extensive classification schemes.

Now, suppose that your protein happens to be a four-stranded β-sheet. That terminology alone is not sufficient to define the structure. There are different *topological* relationships through which the strands could come into contact with each other. Figure 1.20 shows the 12 possible topologies for four-stranded β-sheets, where the third and fourth strands (in sequence) are antiparallel. Some topologies are more common than others.

The next level of description accounts not only for the topological arrangement of secondary structural elements, but for their particular packing geometry in 3D as well. This level of description is called the *fold* of a native protein. A protein fold is a particular arrangement of secondary structures in a tertiary structure. Some common folds are shown in Figure 1.21. The *globin fold* (see Figure 1.21A) has eight α-helices. The jelly roll fold has two superimposed β-sheets (see Figure 1.21B). The *TIM barrel fold* (see Figure 1.21C) takes its name from the enzyme triosephosphate isomerase (TIM), and usually contains eight α-helices (cyan) and eight β-strands (brown). This fold is one of the most common among enzymes. Its successive β–α–β motifs are arranged into a torus. The *Rossmann fold* (see Figure 1.21D) is named after Michael Rossmann, who discovered it in the early 1970s; it binds nucleotides such as nicotinamide adenine dinucleotide [13]. Its fold is β–α–β–α–β.

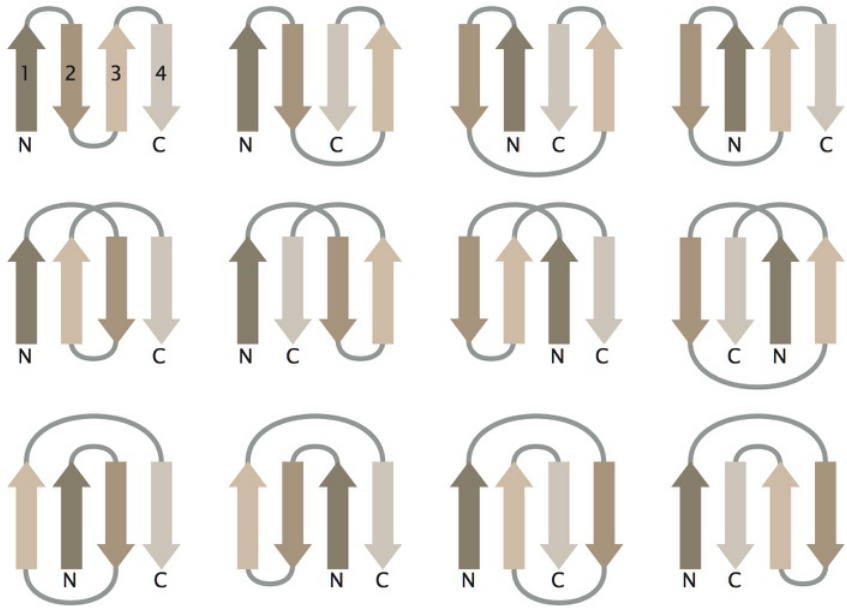

Figure 1.20 Topology diagrams for four-stranded β-sheets. Of all the possible topologies of four-stranded β-sheets, there are only 12 (shown here) that have β-strands three and four in the primary sequence antiparallel to each other and connected by three β–β hairpin loops. Not all topologies in this set are equally probable. Not all of these forms have been observed.

One of the most common protein folds is the β-*barrel*. A β-barrel consists of adjacent β-strands arranged in a cylindrical β-sheet. The cylindrical structure is favored over flat or planar β-sheets because the cylinder leaves no unsatisfied backbone hydrogen bonding groups at the edges. Often the strands are composed of alternating polar and

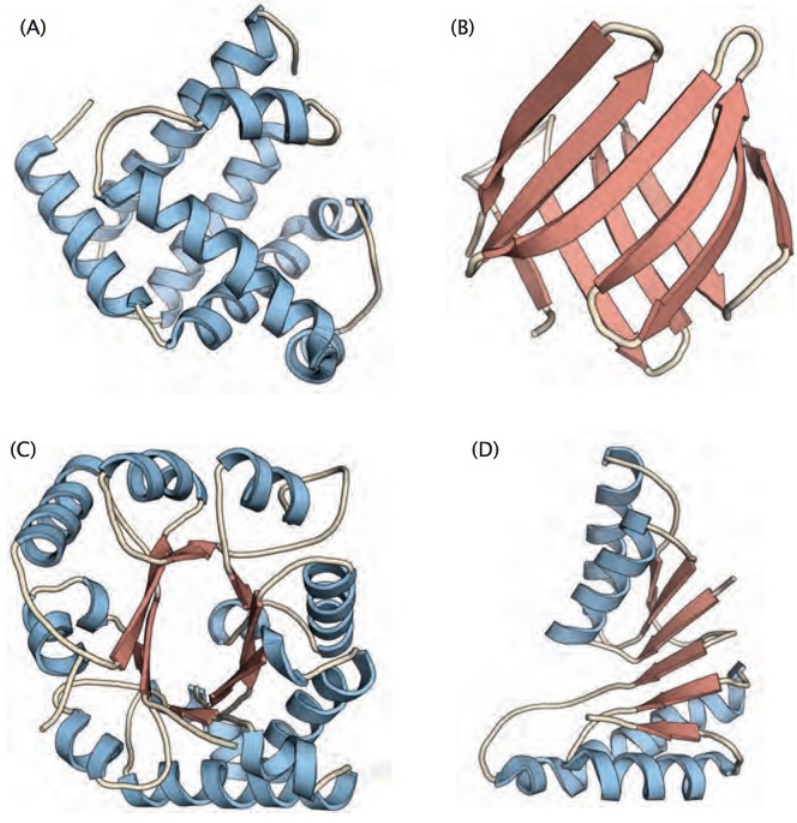

Figure 1.21 Examples of common protein folds. (A) The globin fold contains eight tightly packed α-helices packed together so they can bind an iron-containing heme group (not shown) for storing and transporting oxygen. (B) A jelly roll fold is formed by pairing two antiparallel β sheets together to form a barrel. (C) The left-handed TIM barrel is a doughnut-like shape (toroid) formed by alternating α-helices (*light blue*) and β-strands (*red*) arranged in a closed curve. (D) The Rossmann fold contains two β–α–β–α–β motifs packed to form a central six-stranded β-sheet. In this case, the β-strands (*red*) and α-helices (*blue*) are from a domain of the enzyme decarboxylase, with a flavin mononucleotide cofactor (not shown) bound onto its nucleotide-binding site.

hydrophobic residues. For membrane proteins, such alternation allows for the positioning of the polar groups on the inside, with hydrophobic groups being positioned on the outside of the protein in contact with the membrane lipids. Water-soluble proteins adopt the opposite arrangement: polar groups are on the outside, while hydrophobic groups are inside. The center of the barrel is often a binding site; for example, vitamin A (retinol) binds to the retinol-binding protein.

Proteins Are Classified by Structural and Evolutionary Properties in the CATH Database

A "fold" just describes a geometric property of a protein. You can get additional information for classifying proteins by knowing *evolutionary relationships*, that is, whether or not two proteins evolved from a single ancestor. Such relationships are discussed in more detail in Chapter 7. Here, we just note that there are databases of protein structures that are based on these considerations. Inference of a common evolutionary ancestor is usually based on how similar two amino acid sequences are to each other. A database that classifies proteins based on both their 3D native topologies and their evolutionary relationships to each other is CATH [14]. CATH assigns protein domains into subsets that belong to the same Class, or that have a common Architecture or Topology (fold), or that belong to the same Homologous family (tertiary structure). Figure 1.22 gives the distribution of common CATH classes.

A protein's fold (equivalent to CATH's Topology) is purely geometric. Two proteins having the same fold need not have any particular evolutionary relationship. On the other hand, if two proteins have less than 15% sequence identity while their structures and functions suggest a common evolutionary origin, they may belong to the same *superfamily*. For example, actin, the ATPase domain of the heat-shock protein, and hexokinase belong to the same superfamily. Two proteins are regarded as members of the same *family* if they share 30% or more of their amino acid sequences. If the sequence identity is less than this, down to 15%, proteins may still be in the same family if their functions and 3D structures are very similar. Globins form a family, for example, because they perform the same function, have highly similar structures, and have high sequence identity.

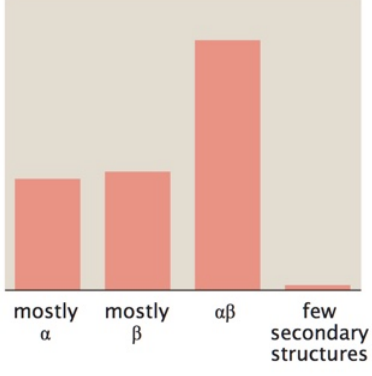

Figure 1.22 Distribution of the protein structures in the PDB in the four CATH classes.

Quaternary Structures: Higher-Order Structures Result from Noncovalent Assemblies of Multiple Chains

In contrast to the domains of tertiary structures, the subunits in a quaternary structure are not covalently bonded to each other, except for occasional disulfide bridge cross-links.

Hemoglobin has a quaternary structure (see Figure 1.11). It carries oxygen in red blood cells. It has four subunits. Each subunit can bind one oxygen molecule, but it is only the full structure of all four subunits that can bind oxygen with sufficient *cooperativity* to pick up oxygen in the lungs and deliver it to the capillaries. Often the symmetries in quaternary structures are important for their function. The cooperativities of ligand binding, for example, depend on the number of ligands that can bind a protein, and that, in turn, depends on how many subunits the protein has. Quaternary structures are called "biological assemblies" in the PDB.

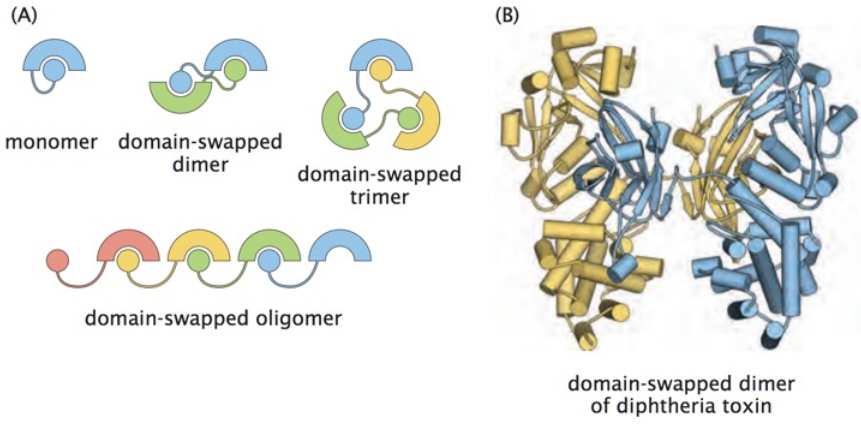

Figure 1.23 Some quaternary structures are constructed by domain swapping. (A) Protein monomers of two domains (arch and ball) are able to form stable quaternary structures ranging from dimer to trimer to oligomer. Stabilizing interactions are contributed at the interfaces formed by swapped domains from different chains. (B) The *yellow* and *blue* chains of the diphtheria toxin dimer swap β domains. (A, adapted from Gronenborn AM. *Curr Opin Struct Biol*, 19:39–49, 2009; B, adapted from MJ Bennett, S Choe, and D Eisenberg. *Proc Natl Acad Sci USA*, 91:3127–3131, 1994.)

Domain Swapping Is Another Way that Proteins Can Form Quaternary Structures

In some cases, quaternary structures are defined by the interdigitation of two or more proteins. In *domain swapping*, one domain (or secondary structure) of a monomer replaces the same domain (or secondary structure) from a different, identical monomer, giving rise to an intertwined dimer or oligomer (Figure 1.23). Sometimes, domain swapping can repeat, chaining together one protein to the next, leading to an ordered form of protein assembly. Fibronectin modules form fibrils from such chains of domain-swapped proteins.

SOME PROTEINS ARE STABLE AND FUNCTION IN THE MEMBRANE ENVIRONMENT

Membrane proteins are localized in the membranes of cells or organelles, such as mitochondria. Some membrane proteins are channels that allow the flow of ions, such as potassium. Others function as electron or proton pumps (for example, cytochrome *c* oxidase and complex IV in mitochondria, and ATPase in cell membranes), receptors (for example, G-protein-coupled receptors, GPCRs) or transporters (for example, glutamate transporters and ABC transporters) across the membrane (Figure 1.24). Some membrane proteins can function as receptors and ion channels at the same time (AMPA and NMDA). These are all *integral membrane proteins*. They are stable and functional when embedded in the lipid bilayer. They play a key role in maintaining or regulating the physiological levels of ions and substrates at the extracellular and intracellular regions, assisting with establishing appropriate concentrations or energy gradients across the membrane, and enabling signal transduction events across the membrane. *Peripheral membrane proteins* are temporarily attached (usually from the extracellular side) to cell membranes or to integral membrane proteins. Here we are referring to integral membrane proteins, unless otherwise stated. Membrane proteins are often important drug targets.

A membrane protein usually has three regions: transmembrane (TM), extracellular, and intracellular. The membrane is typically a lipid bilayer, which is a sandwich of two layers of lipid molecules. The polar

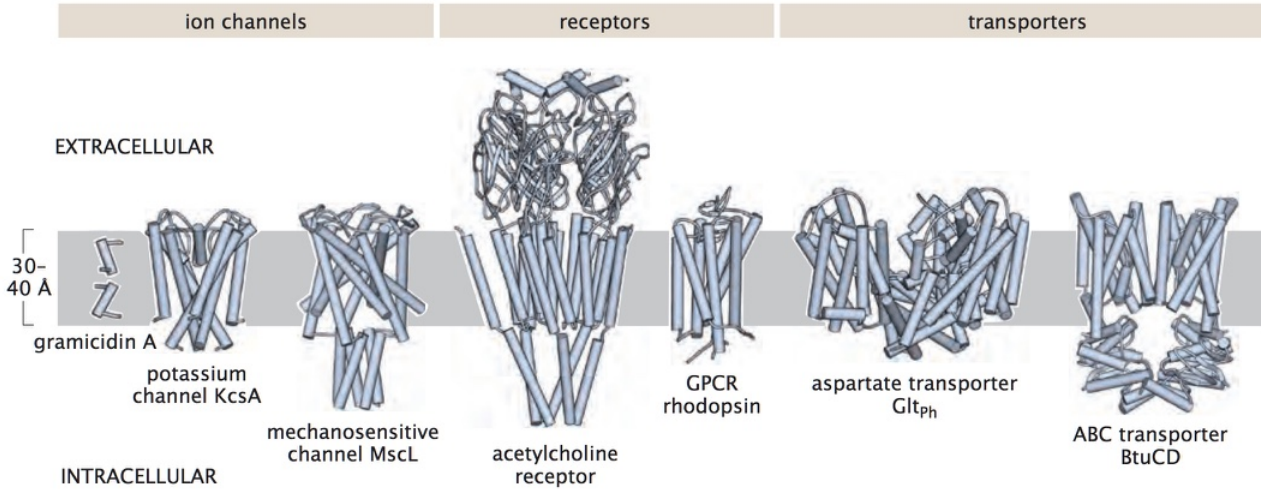

Figure 1.24 Different sizes and types of membrane proteins. The sizes range from small helical peptides in gramicidin A to multimeric proteins with large extracellular and/or intracellular domains, as in the acetylcholine receptor. (Adapted from I Bahar, TR Lezon, A Bakan, and IH Shrivastava. *Chem Rev*, 110:1463–1497, 2010.)

heads groups face outward toward the water and the hydrophobic tails face each other inside the bilayer. In contrast to globular proteins, which have hydrophobic groups buried in the core and polar/charged groups facing the surrounding water, the surfaces of membrane proteins usually contain hydrophobic residues that make favorable contacts with the surrounding lipid molecules inside the membrane.

An important class of membrane proteins is the GPCRs (G-protein-coupled receptors). GPCRs have a seven-TM-helix fold. β-barrel folds are also commonly observed, for example in porin proteins. Of the 1700 TM structures known in 2012, about 1400 are α-helical and 250 are β-barrels. Many membrane proteins are multimeric. The monomers assemble to form a central pore (for example, potassium channels, which are tetramers), or a stable scaffold that supports the cooperative transition between outward-facing and inward-facing forms (for example, glutamate transporters, which are homotrimers, and the ABC transporters, which are heterodimers).

SOME PROTEINS HAVE FIBROUS STRUCTURES

Fibrous proteins are elongated, with a single dominant type of secondary structure. For example, collagen forms a triple-helical right-handed coiled coil, so it has great mechanical strength. Fibrous protein sequences are often highly repetitive. Collagen has long stretches of repeats of the tripeptide Gly-Pro-X, where X can be any amino acid (Figure 1.25). A large fraction of the world's protein mass is fibrous. Collagen, which is the main protein of connective tissue, is the most abundant protein in vertebrates, making up 25–35% of their whole-body protein content.

The mechanical and load-bearing frameworks of organisms are constructed from fibrous proteins. Collagen forms the stress-bearing elements of skin, bone, teeth, and tendons. β-*keratin* is a two-helix coiled coil, found in fur and claws. The essential protein in silk is called *fibroin*; it forms a β-sheet. Some fibrous proteins are elastic. For

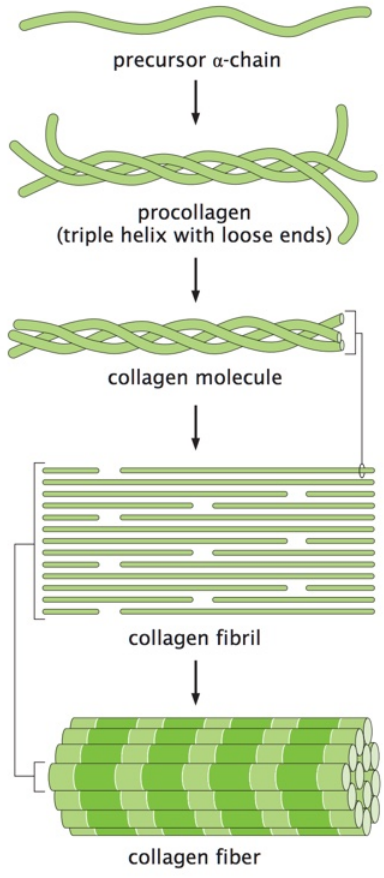

Figure 1.25 Fibrous proteins are made from α-helical coiled coils or stacked β-sheets. They are mechanically strong and can extend to macroscopic sizes. The collagen triple helix is a right-handed superhelix or coiled coil. The three helices adopt polyproline II conformations and twist around their neighbors like strands in a rope. The stability is maintained by extensive hydrogen bonding between the neighboring helices.

example, elastin, found in ligaments and the walls of the lungs and blood vessels, is primarily composed of small, nonpolar amino acids (Ala, Val, Gly), Pro, and Lys, and acts as a collection of springs that are cross-linked into an irregular assembly.

Fibrous Proteins Include Coiled Coils and β-Helices

Coiled coils occur when at least two α-helices are wound around each other in a regular twist, like the strands of a rope (see Figure 1.25). They are among the most common supersecondary structures, and can contain two, three, four, or five helices. The component helices may be aligned either antiparallel or parallel. Figure 1.25 shows a coiled coil of three helices. Coiled coils are ubiquitous, occurring in transcription factors, viral fusion peptides, and certain tRNA synthetases.

Other fibrous structures are based on β-strands. Silk fibroin has a regular amino acid sequence repeat of -(Gly-Ser-Gly-Ala-Gly)-. It has high mechanical strength because of its extensive β-structure. Silk is strong along the fiber axis because that is the direction of the chain's covalent bonding. But in the other directions, silk is more flexible because

the chains are bonded only by hydrogen bonds, which are weaker than covalent bonds.

NATIVE PROTEINS ARE CONFORMATIONAL ENSEMBLES

Proteins Fluctuate around their Native Structures

So far, we have focused on single native structures of proteins. However, that is not the whole story. First, even well-defined native protein structures wiggle and shake at room temperature because they are bombarded by the Brownian motion of the surrounding solvent molecules. NMR methods can reveal the fluctuations of a protein around its native structure. Multiple chain conformations are consistent with the NMR determination of the structure of a protein, indicating the fluctuations around a native protein structure. When you refer to "the" native structure, you are referring to a representative structure from this ensemble.

A Protein Can Sample Multiple Substates under Native Conditions

Second, proteins can deviate from the "single-native-structure" ideal in another respect. They can undergo conformational changes. For example, the structure of a protein in the absence of a bound ligand is called its *apo* state, while the structure in the presence of a bound ligand is called its *holo* state. Figure 1.26 shows adenylate kinase, which undergoes changes in conformation upon ligand binding, changing from an "open" to a "closed" conformation, with ligand-binding stabilizing the closed form.

Another commonly observed structural transition, in the case of membrane proteins, is the passage from the so-called outward-facing conformation to the inward-facing conformation, known as the *alternating access model.* The outward-facing conformation is open to the extracellular environment to bind/uptake the ligand, and the inward-facing conformation releases the substrate to the intracellular region. Transitions between these two functional structures are essential to the transport of substrates by transporters (membrane proteins) across the cell membrane.

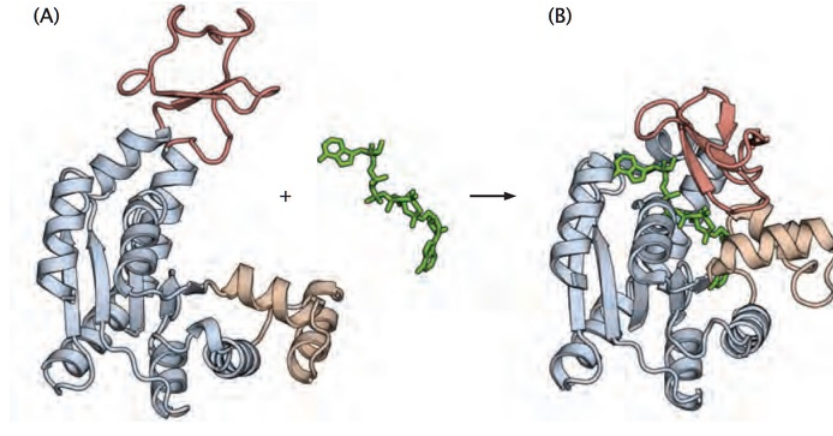

Figure 1.26 Adenylate kinase (colored by its three domains) is an example of a protein that exists in an open form (A) and a closed form (B). As in many cases, the ligand (a nucleotide, *green*) binds and the protein then closes around it. (Adapted from U Olsson and M Wolf-Watz. *Nat Commun*, 1:111, 2010. With permission from Macmillan Publishers Ltd.)

Some Proteins Are Intrinsically Disordered

A third way that proteins can deviate from the "single-native structure" paradigm is through intrinsic disorder. Some proteins have regions that are *intrinsically disordered*, meaning that those parts of the chain are not well defined in experimental structure determinations. Interestingly, intrinsic disorder can serve functional purposes. For example, a positively charged intrinsically disordered protein can bind to a negatively charged DNA molecule, causing the complex to form a unique structure upon binding, a phenomenon referred to as *folding upon binding*.

In another form of disorder, proteins can form *molten globule* states: relatively compact structures having residual native-like secondary structures but little tertiary structure. Proteins sometimes form molten globular states, for example, under acidic pH conditions.

SUMMARY

Proteins are polymer chains. They fold into compact states that are diverse in size, shape, and dynamics. They perform many different biological functions. Different proteins have different sequences of the 20 types of building-block amino acids. Some amino acids are nonpolar, some are polar, and some are positively or negatively charged. Different amino acid sequences fold into different 3D shapes. Folded proteins adopt structures on different levels: secondary structures include helices and sheets; tertiary structures are assemblies of secondary structures in well-defined folds; and quaternary structures are composed of multiple chains (or subunits). Proteins may be globular and soluble in water, or fibrous, or may be localized within membrane environments. Chapter 2 gives an overview of how a protein's biological actions are encoded in its 3D native structure and motions.

REFERENCES

[1] R Milo. What is the total number of protein molecules per cell volume? A call to rethink some published values. *Bioessays*, 35:1050–1055, 2013.

[2] HM Berman, J Westbrook, Z Feng, et al. The protein data bank. *Nucleic Acids Res*, 28:235–242, 2000.

[3] GN Ramachandran, C Ramakrishnan, and V Sasisekharan. Stereochemistry of polypeptide chain configurations. *J Mol Biol*, 7:95–99, 1963.

[4] JA Rupley and G Careri. Protein hydration and function. *Adv Protein Chem*, 41:37–172, 1991.

[5] K Linderstrom-Lang. Proteins and enzymes. *Lane Medical Lectures*, Stanford University Publications, University Series, Medical Sciences, VI:1–115, 1952.

[6] L Pauling, RB Corey, and HR Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA*, 37:205–211, 1951.

[7] H Tadokoro. *Structure of Crystalline Polymers*. Wiley, New York, 1979.

[8] L Pauling and RB Corey. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. *Proc Natl Acad Sci USA*, 37:729–740, 1951.

[9] S Kanamaru, PG Leiman, VA Kostyuchenko, et al. Structure of the cell-puncturing device of bacteriophage T4. *Nature*, 415:553–557, 2002.

[10] MJE Sternberg and JM Thornton. On the conformation of proteins: The handedness of the connection between parallel β-strands. *J Mol Biol*, 110:269–283, 1977.

[11] TWF Slidel and JM Thornton. Chirality in protein structure. In H Bohr and S Brunak, editors, *Protein Folds: A Distance Based Approach*, pp 253–264. CRC, Boca Raton, FL, 1995.

[12] M Levitt and C Chothia. Structural patterns in globular proteins. *Nature*, 261:552–558, 1976.

[13] ST Rao and MG Rossmann. Comparison of super-secondary structures in proteins. *J Mol Biol*, 76:241–250, 1973.

[14] F Pearl, A Todd, I Sillitoe, et al. The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acid Res*, 33:D247–D251, 2005.

SUGGESTED READING

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

Kuriyan J and Konforti B, *The Molecules of Life: Physical and Chemical Principles*. Garland Science, New York, 2012.

Voet D and Voet JG, *Biochemistry*, 4th ed. Wiley, Hoboken, NJ, 2011.