

# CHAPTER 4

## Protein Structure

This chapter is concerned with the architectural principles underlying the three-dimensional structures of protein molecules. We shall first discuss some general principles, including the hierarchical organization of protein structure and the role of the hydrophobic effect and hydrogen bonding in stabilizing it. We then study the nature of conformational restrictions on the peptide backbone. These restrictions arise from collisions between backbone atoms in certain conformations, which are therefore disallowed. We shall see how the restricted set of allowed conformations, when combined with hydrogen bonding, lead naturally to the secondary structural elements known as the  $\alpha$  helix and the  $\beta$  strand as the fundamental building blocks of protein structure.

There are two major classes of proteins, called water-soluble proteins and membrane proteins, which are categorized based on whether or not they interact with cell membranes. Water-soluble proteins include globular proteins, which have compact and well-defined structures, and are not associated with membranes. Globular proteins have mostly hydrophobic sidechains on the inside and mostly hydrophilic sidechains on the outside (see Figure 1.35). In this chapter, the only soluble proteins we shall focus on are the globular proteins. We shall not discuss fibrous proteins, such as collagen. Nor do we study intrinsically disordered proteins—a class of proteins that do not have well-defined structure on their own.

Membrane proteins, which are discussed in this chapter, are embedded either partially or completely in the interior of the hydrophobic part of a lipid bilayer, and they have hydrophobic sidechains on the outside as well as on the inside. The principles underlying the structures of membrane proteins are a little different than those of globular proteins, but both consist of  $\alpha$  helices and  $\beta$  strands. We shall first survey the principles governing the folded structures of globular proteins and describe some of the commonly occurring kinds of protein structures. These principles will help us understand how the architectures of membrane proteins, discussed at the end of the chapter, differ from those of globular proteins. Membrane proteins are particularly important for the generation of energy and for transporting molecules into and out of the cell, and we discuss two examples of proteins that carry out these functions.

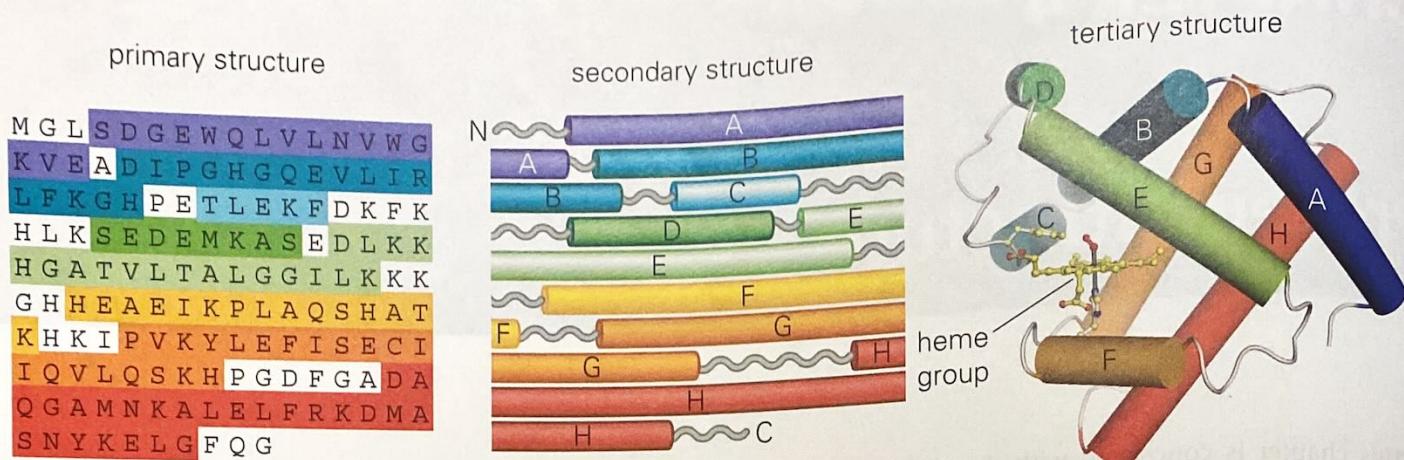
### A. GENERAL PRINCIPLES

#### 4.1 Protein structures display a hierarchical organization

Recall from Chapter 1 that the folded structures of protein molecules are built up from the packing together of two kinds of secondary structure elements:  $\alpha$  helices and  $\beta$  strands. The terms “primary,” “secondary,” and “tertiary” structure are used to emphasize the hierarchical levels in the structures of proteins. **Primary structure** is the amino acid sequence—that is, the pattern of amino acids along a linear polypeptide chain (Figure 4.1). The local conformation of the polypeptide chain is referred to as the **secondary structure**, which occurs mainly as  $\alpha$  helices and  $\beta$  strands, with loops connecting them.

#### Primary, secondary, tertiary, and quaternary structure

The amino acid sequence of a protein is referred to as its primary structure. The local conformation of the protein backbone ( $\alpha$  helix,  $\beta$  strand, or loop) is the secondary structure. The three-dimensional fold of a protein is the tertiary structure. Finally, the arrangement of subunits in a multi-subunit protein complex is the quaternary structure.



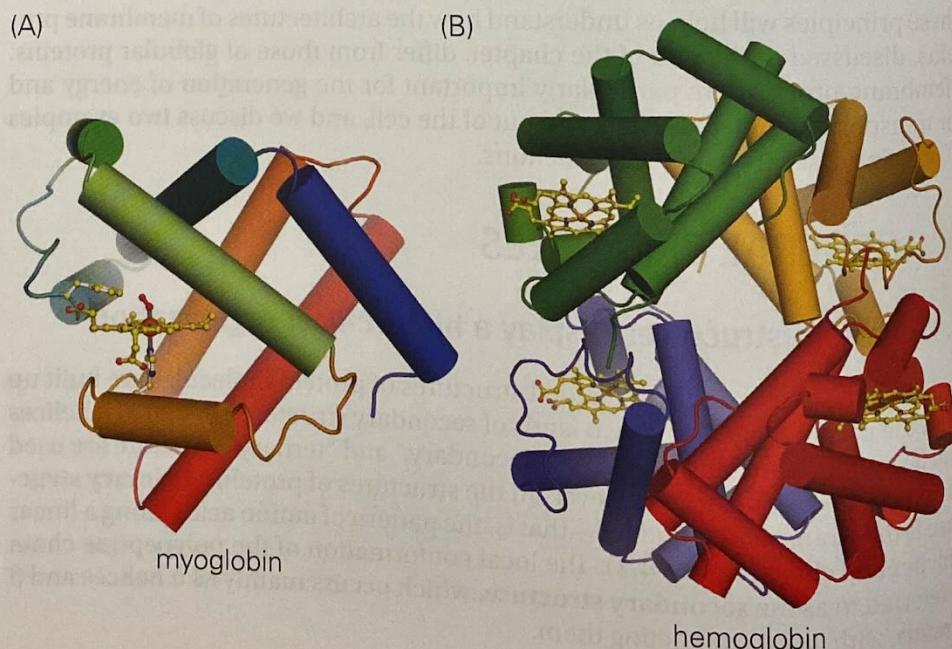
**Figure 4.1 Primary, secondary, and tertiary structure.** The amino acid sequence of human myoglobin, an oxygen-binding protein, is shown on the left. The secondary structure of myoglobin is shown in the middle. The  $\alpha$  helices are indicated as cylinders and the loops connecting them as wavy lines (myoglobin contains no  $\beta$  strands). The tertiary structure of myoglobin is shown on the right. The particular three-dimensional arrangement of  $\alpha$  helices that is characteristic of myoglobin is known as the globin fold and is also seen in hemoglobin (see Figure 4.2). Oxygen binds to an iron atom that is part of a cofactor known as a heme group (yellow bonds), which is discussed in more detail in Chapters 5 and 14. The heme group is not part of the protein chain. The sidechains of two amino acid residues that interact with the heme group are also shown. (PDB code: 1MBC.)

### Structural domains

Structural domains are the fundamental units of three-dimensional protein structure. A protein domain is typically 50–200 residues long and contains a well-defined hydrophobic core. Domains are mixed and matched in evolution to produce more complex proteins.

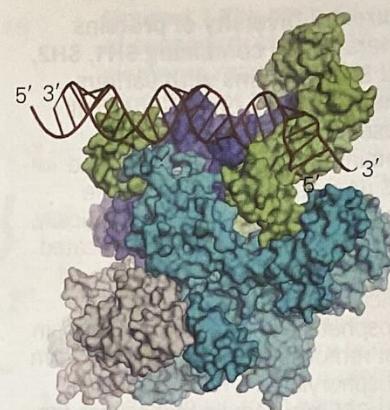
As a general rule, the residues that form the interfaces between secondary structural elements are hydrophobic, and the packing of secondary structural elements results in the formation of a protein **structural domain** with a hydrophobic core. Protein molecules contain one or more such domains. **Tertiary structure** refers to the three-dimensional organization of the secondary structure elements in these domains. The particular three-dimensional arrangement of the  $\alpha$  helices and  $\beta$  strands in the tertiary structure is known as the **protein fold**.

Many proteins consist of several different polypeptide chains (subunits) that associate into a multimeric complex in a specific way called the **quaternary structure**. These subunits can function either independently of each other or cooperatively so that the function of one subunit depends on the functional states of the other subunits. The quaternary structure of the oxygen transport protein hemoglobin is illustrated in Figure 4.2. Hemoglobin consists of four subunits that are structurally similar to each other. In other protein complexes, such as RNA polymerase (Figure 4.3), the subunits are structurally distinct.



**Figure 4.2 Quaternary structure of human hemoglobin.** Hemoglobin, the oxygen transport protein, is made up of two copies each of two different subunits, both of which look like myoglobin. (A) The tertiary structure of myoglobin, illustrating the globin fold. (B) The quaternary structure of hemoglobin. (PDB codes: 1MBC and 1A00.)

**Figure 4.3 Quaternary structure of bacterial RNA polymerase.** This enzyme, which transcribes genetic information from DNA into RNA, is composed of several different kinds of subunits, each of which is shown in a different color. A DNA molecule bound to RNA polymerase is also shown. (Adapted from K.S. Murakami et al., and S.A. Darst, *Science* 296: 1285–1290, 2002. With permission from the AAAS; PDB code: 1L9Z.)



## 4.2 Protein domains are the fundamental units of tertiary structure

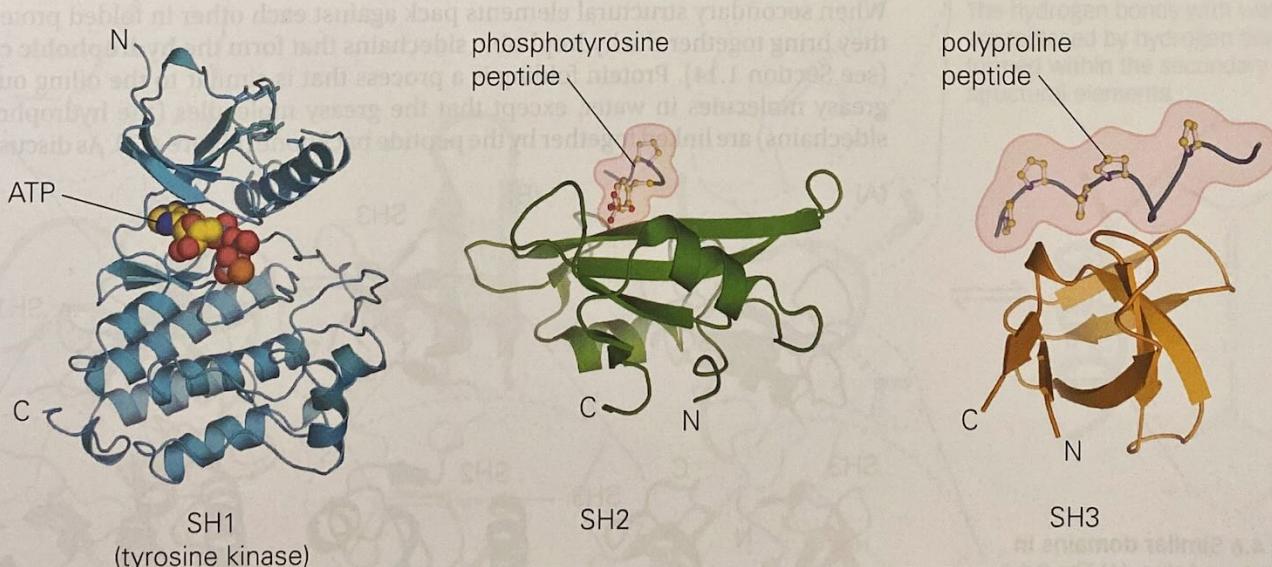
A protein structural domain is a polypeptide chain or a part of a polypeptide chain that forms an independent structural unit, typically with a well-defined hydrophobic core. Domains are often units of function, and proteins may comprise a single domain or several. Protein domains range in size from ~50 to ~200 amino acid residues and represent a great many, perhaps thousands, of distinct polypeptide chain folds.

Domains are the fundamental building blocks in the evolution of complex protein structure, as discussed in Chapter 5 (Part D). Gene duplication and genetic recombination have resulted in the combinatorial creation of multifunctional proteins by mixing and matching structural domains. We illustrate this idea by using three particular domains as examples. These domains are important in the processes that control signaling between cells, and are known as the Src homology (SH) domains—namely, the SH1, SH2, and SH3 domains (**Figure 4.4**). These names reflect the original identification of these domains in a family of proteins that are related to (that is, are *homologous* to) a protein known as Src (pronounced *sark*). The Src protein is so named because it causes a type of cancer known as a sarcoma (hence the abbreviation, *Src*, for sarcoma).

SH1 domains are enzymes known as tyrosine kinases, and they catalyze the transfer of phosphate groups from ATP to tyrosine residues on other proteins. SH2 domains bind to phosphorylated tyrosine residues in other proteins, and SH3 domains bind to peptide segments containing proline residues at specific positions. There are dozens of distinct SH1, SH2, or SH3 domains in a mammalian genome. Individual SH1, SH2, and SH3 domains are combined with each other or with a variety of other domains to generate a diversity of proteins with specialized

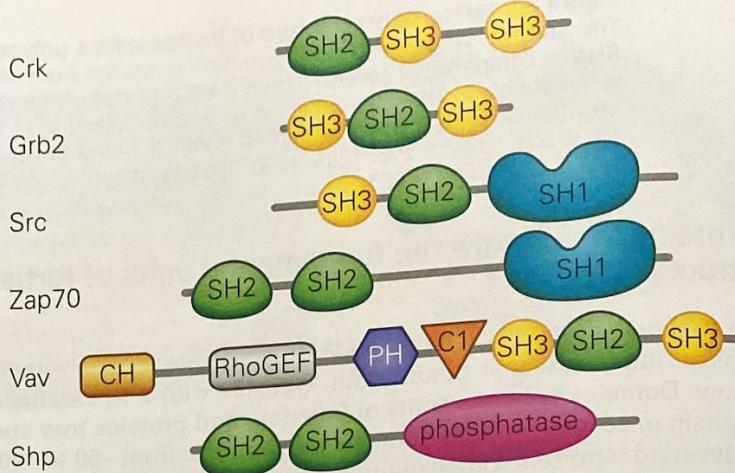
### Protein fold

The visually recognizable arrangement of  $\alpha$  helices and  $\beta$  strands in the three-dimensional structure of a protein is referred to as the protein fold. Different proteins can have the same fold, as is the case for myoglobin and the subunits of hemoglobin (see Figure 4.2).



**Figure 4.4 The Src homology domains SH1, SH2, and SH3.** The SH1 domain (blue) is a tyrosine kinase enzyme and is shown here bound to ATP. Tyrosine kinases catalyze the transfer of the terminal phosphate of ATP to proteins that they bind to. SH2 domains (green) bind to segments of proteins that contain a phosphorylated tyrosine residue (shown shaded in red). SH3 domains (yellow) bind to protein segments containing proline residues (shaded red). [PDB codes: 1M14 (SH1 domain), 1SHA (SH2 domain), and 1SRL (SH3 domain).]

**Figure 4.5 Diversity of proteins generated by combining SH1, SH2, and SH3 domains with various other domains.** Only a very small subset of the large number of proteins containing SH1, SH2, and SH3 domains is shown here. The proteins are illustrated schematically, with the polypeptide chain indicated by a line. Different domains are indicated by different symbols. Phosphatase is an enzymatic domain that removes phosphate groups from phosphorylated tyrosine residues. Rho-GEF is a domain that causes the dissociation of GTP or GDP bound to proteins of the Rho family. CH domains bind to the actin cytoskeleton, and PH and C1 domains bind to specific lipids in the cell membrane. (Adapted from T. Pawson, M. Raina, and P. Nash, *FEBS Lett.* 513: 2–10, 2002. With permission from Elsevier.)

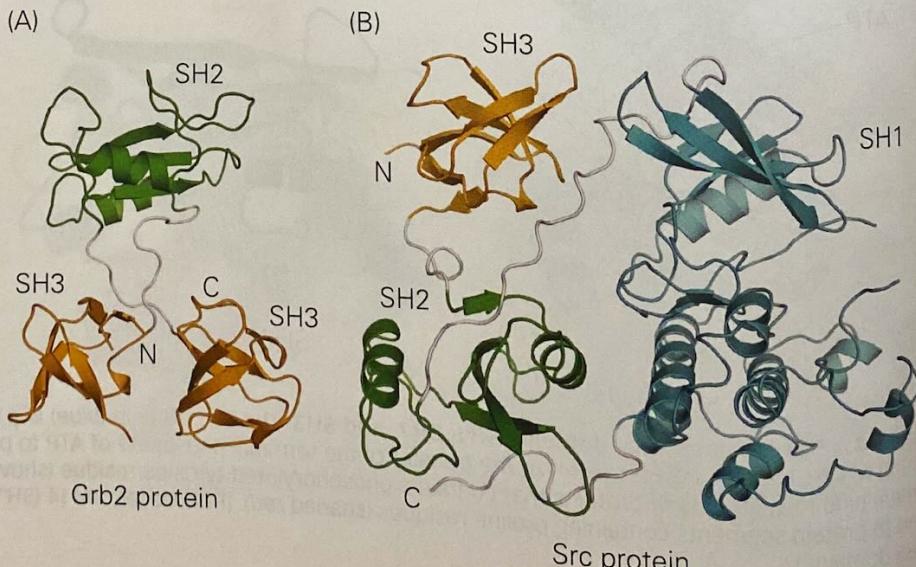


functions (Figure 4.5). Notice that while the Src protein has an SH2 domain, an SH3 domain, and an SH1 domain connected in that order, some of the other proteins illustrated in Figure 4.5 lack one or two of these domains, have multiple copies of one kind of domain, or have other kinds of domains included in their structures.

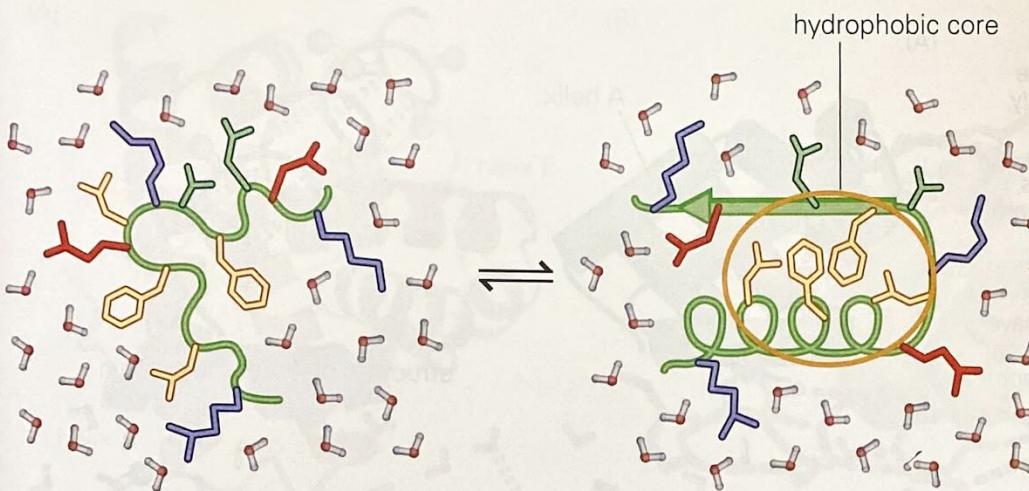
The SH1, SH2, or SH3 domains in the proteins shown in Figure 4.5 have different amino acid sequences. As we discuss in the next chapter, changes in the sequences of the domains, during evolution, result in slightly different structures and in altered specificities for the proteins that they interact with. Although the general fold of each of the domains is conserved, the three-dimensional arrangement of these domains in different proteins can be quite different. This is illustrated in Figure 4.6, which compares the structure of a protein that contains an SH2 domain and two SH3 domains, known as Grb2, and the Src protein. Note, as shown in Figure 4.6, that the SH2 and SH3 domains are arranged quite differently in Grb2 and Src.

### 4.3 Protein folding is driven by the formation of a hydrophobic core

When secondary structural elements pack against each other in folded proteins, they bring together the hydrophobic sidechains that form the **hydrophobic core** (see Section 1.14). Protein folding is a process that is similar to the oiling out of greasy molecules in water, except that the greasy molecules (the hydrophobic sidechains) are linked together by the peptide backbone (Figure 4.7). As discussed



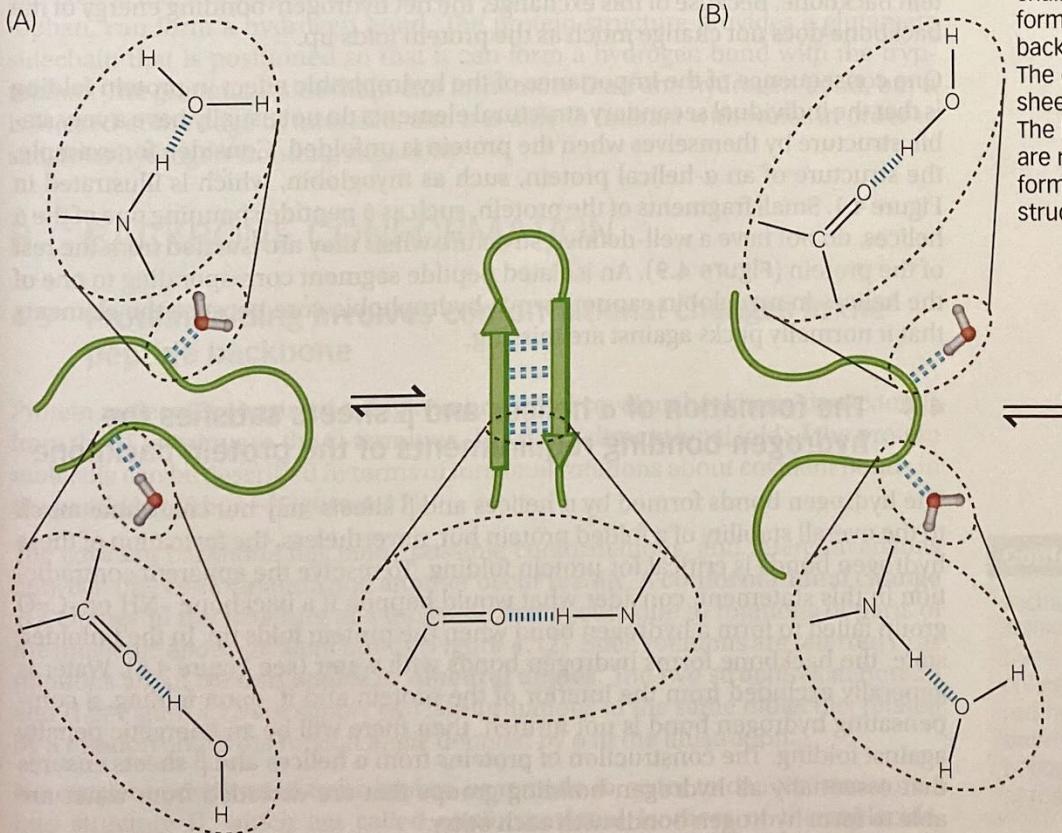
**Figure 4.6 Similar domains in different proteins.** (A) The Grb2 protein has an SH2 domain and two SH3 domains. (B) The Src protein has an SH2 domain and an SH3 domain attached to an SH1 domain. (PDB codes: 1GRI and 2HCK.)



**Figure 4.7** The packing of secondary structural elements results in the formation of the hydrophobic core. The backbone of a small protein molecule is shown schematically here, along with a few of the sidechains. In the unfolded state, all of the sidechains are exposed to water. When the protein folds up, most of the hydrophobic sidechains (yellow) are buried in the interior of the protein and are removed from water. The charged and polar sidechains remain exposed to water.

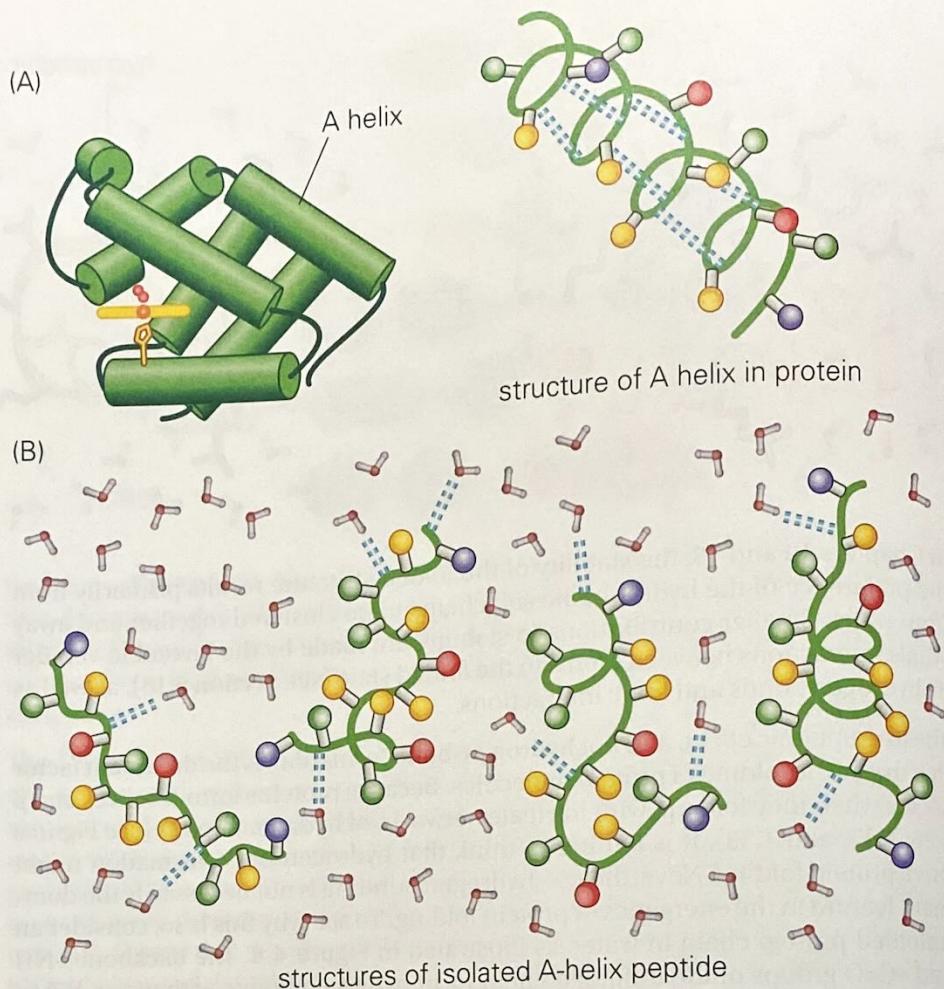
in Chapters 10 and 18, the stability of the folded structure results primarily from the preference of the hydrophobic sidechains to be clustered together and away from water. Smaller contributions to stability are made by the favorable van der Waals interactions between atoms in the folded state (see Section 6.16), as well as by hydrogen bonds and ionic interactions.

The hydrophobic effect, and not hydrogen-bond formation, is the dominant factor that drives the folding of protein molecules. Because proteins form  $\alpha$  helices and  $\beta$  sheets when they fold up, with intricate networks of hydrogen bonds (see Figures 1.36, 1.37, and 1.38), it is natural to think that hydrogen-bond formation might drive protein folding. Nevertheless, hydrogen bonding is not necessarily the dominant feature in the energetics of protein folding. To see why this is so, consider an unfolded protein chain in water, as illustrated in **Figure 4.8**. The backbone  $-NH$  and  $-C=O$  groups of the unfolded chain form hydrogen bonds with water. When the chain folds up, the formation of  $\alpha$  helices and  $\beta$  sheets results in the hydrogen



**Figure 4.8** Exchange of hydrogen bonds between backbone groups and water molecules. The left-hand side of each panel in this figure shows an unfolded protein chain and how water molecules form hydrogen bonds with the backbone  $-NH$  and  $-C=O$  groups. The chains fold up to form a  $\beta$  sheet in (A) and an  $\alpha$  helix in (B). The hydrogen bonds with water are replaced by hydrogen bonds formed within the secondary structural elements.

**Figure 4.9 Isolated fragments of proteins do not usually have stable structures.** (A) The structure of myoglobin is shown schematically, with the A helix indicated. The A helix is also shown separately, with the sidechains indicated by spheres, colored according to whether they are hydrophobic (yellow), polar (green), or charged (red or blue). (B) A short peptide with the same sequence as the A helix of myoglobin does not have a stable structure in solution, although it shows a tendency to form transient helical turns.

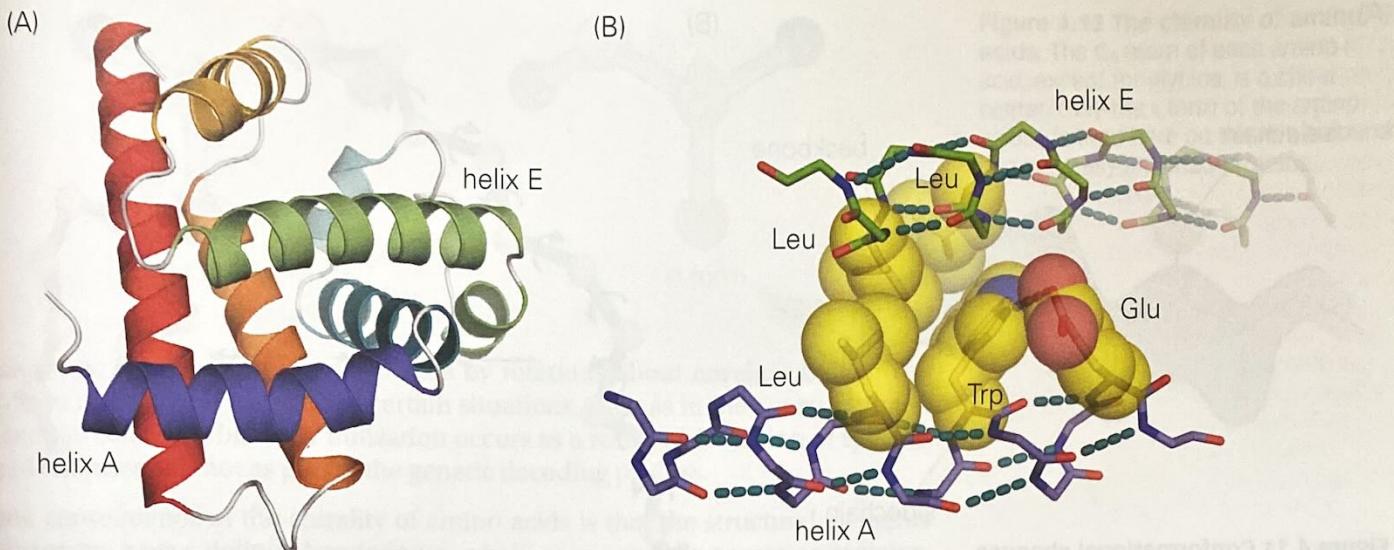


bonds with water being replaced by hydrogen bonds with other parts of the protein backbone. Because of this exchange, the net hydrogen-bonding energy of the backbone does not change much as the protein folds up.

One consequence of the importance of the hydrophobic effect in protein folding is that the individual secondary structural elements do not usually have a very stable structure by themselves when the protein is unfolded. Consider, for example, the structure of an  $\alpha$ -helical protein, such as myoglobin, which is illustrated in Figure 4.1. Small fragments of the protein, such as a peptide spanning one of the  $\alpha$  helices, do not have a well-defined structure when they are isolated from the rest of the protein (Figure 4.9). An isolated peptide segment corresponding to one of the helices in myoglobin cannot form a hydrophobic core because the elements that it normally packs against are missing.

#### 4.4 The formation of $\alpha$ helices and $\beta$ sheets satisfies the hydrogen-bonding requirements of the protein backbone

The hydrogen bonds formed by  $\alpha$  helices and  $\beta$  sheets may not contribute much to the overall stability of a folded protein but, nevertheless, the formation of these hydrogen bonds is critical for protein folding. To resolve the apparent contradiction in this statement, consider what would happen if a backbone  $-NH$  or  $-C=O$  group failed to form a hydrogen bond when the protein folds up. In the unfolded state, the backbone forms hydrogen bonds with water (see Figure 4.8). Water is generally excluded from the interior of the protein and if, upon folding, a compensating hydrogen bond is not formed, then there will be an energetic penalty against folding. The construction of proteins from  $\alpha$  helices and  $\beta$  sheets ensures that essentially all hydrogen-bonding groups that are excluded from water are able to form hydrogen bonds with each other.



**Figure 4.10 Folded protein structures satisfy hydrogen-bonding requirements.** (A) The structure of myoglobin is shown. Two  $\alpha$  helices, denoted A and E, pack together in the folded structure. An expanded view of the two helices is shown in (B), illustrating how hydrophobic sidechains (that is, three leucine residues and one tryptophan residue) are brought close together by the packing of the helices. Hydrogen bonds are indicated by dashed lines. Notice that the hydrogen-bonding capability of the backbone  $-NH$  and  $-C=O$  groups are satisfied by the formation of the two  $\alpha$  helices. The tryptophan sidechain has a nitrogen atom that can form a hydrogen bond, and a glutamate sidechain is positioned so that it can form a hydrogen bond with the tryptophan. The glutamate sidechain is exposed to water, with which it also forms hydrogen bonds (not shown here). (PDB code: 1A6M.)

This principle is illustrated in **Figure 4.10**, which shows how two  $\alpha$  helices pack together in myoglobin. The backbone groups form hydrogen bonds in a regular pattern, thereby compensating for the loss of hydrogen bonds with water, which is excluded from the interface between the two helices. Water molecules can sometimes be trapped within the folded structure, but this is relatively rare. The sidechains that are located at the interface are hydrophobic. One of these, a tryptophan, can form a hydrogen bond. The protein structure provides a glutamate sidechain that is positioned so that it can form a hydrogen bond with the tryptophan. The glutamate sidechain can form more than one hydrogen bond, but it is located at the edge of interface, and it is able to interact with water in order to satisfy its hydrogen-bonding capability.

## B. BACKBONE CONFORMATION

### 4.5 Protein folding involves conformational changes in the peptide backbone

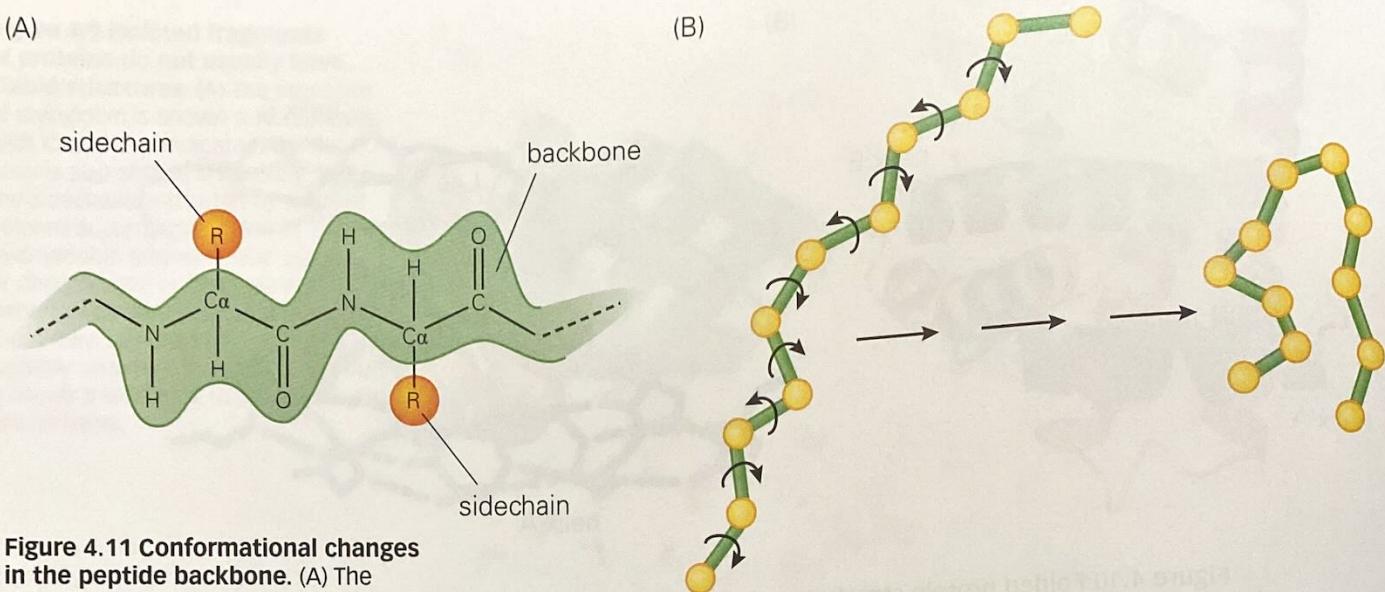
Protein molecules consist of a single unbranched covalent backbone that extends from the N-terminus to the C-terminus. The three-dimensional fold of the protein molecule can be described in terms of torsional rotations about covalent bonds in the peptide backbone (**Figure 4.11**).

The covalent backbone has many possible conformations, and interconversions between one conformation and another occur readily. A **conformational change** is a change in the structure of a molecule that occurs due to rotations of parts of the molecule about covalent bonds (**Figure 4.12**). Such rotations are referred to as rotations about **torsion angles** or **dihedral angles**. The two structures denoted I and II in Figure 4.12 are simply two conformations of the same molecule, related by a rotation about the torsion angle denoted by  $\phi$  in the illustration.

No covalent bonds need to be broken or remade in order to convert structure I into structure II, which are called **conformational isomers** of the molecule.

#### Conformational change

A change in structure that arises solely from rotations about covalent bonds is called a conformational change. Conformational rearrangements do not involve breaking and forming covalent bonds, and can often occur readily at room temperature.

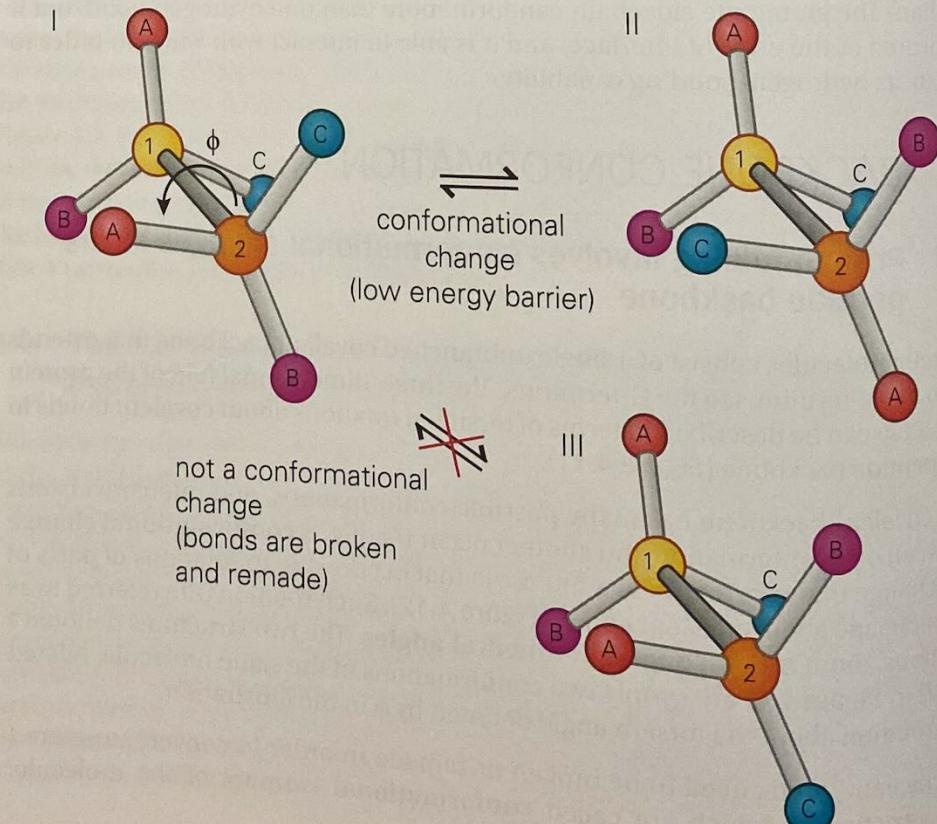


**Figure 4.11 Conformational changes in the peptide backbone.** (A) The chemical structure of a polypeptide (protein) chain, indicating the backbone. (B) A schematic diagram showing the folding of a protein chain into a compact three-dimensional structure. Each amino acid residue is represented by a sphere. The folding process is primarily a conformational change in the protein backbone.

Consider, instead, the structure denoted III in Figure 4.12. This structure cannot be obtained from the other two structures without breaking and remaking covalent bonds. Structure III is a **stereoisomer** of the other two structures.

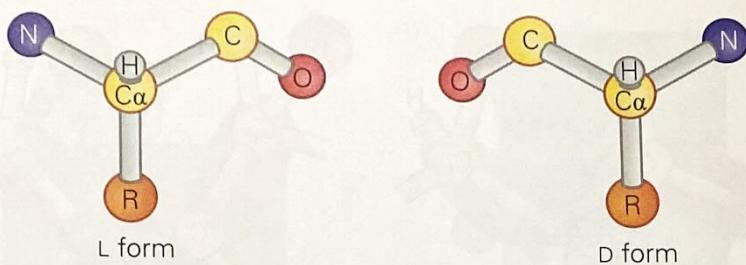
#### 4.6 Amino acids are chiral and only the L form stereoisomer is found in genetically encoded proteins

Molecules that form stereoisomers have special atoms known as **chiral centers**, which are bonded to inequivalent groups of atoms. The  $C_\alpha$  atom of each amino acid, except for glycine, is a chiral center because it is bonded to the N atom of the amide, the C atom of the carbonyl, a hydrogen atom, and the R group (Figure 4.13). Amino acids, except for glycine, have two stereoisomers, known as the **L form** and the **D form**, which are illustrated in Figure 4.13. Only the L forms of the amino acids are found in genetically encoded proteins. Note that L amino acids



**Figure 4.12 Conformational changes and stereoisomers.**

A simple molecule is illustrated here. Two of its atoms, labeled 1 and 2 are bonded together, and each has three substituents, labeled A, B, and C. A conformational change would simply rotate one set of substituents with respect to the other. Any conversion that requires making or breaking bonds is not a conformational change. For example, the molecular structure labeled III, which is a stereoisomer of molecule I, cannot be obtained from the structure labeled I without breaking and remaking covalent bonds.



**Figure 4.13 The chirality of amino acids.** The  $C_{\alpha}$  atom of each amino acid, except for glycine, is a chiral center. Only the L form of the amino acids, defined here on the left, is found in genetically encoded proteins.

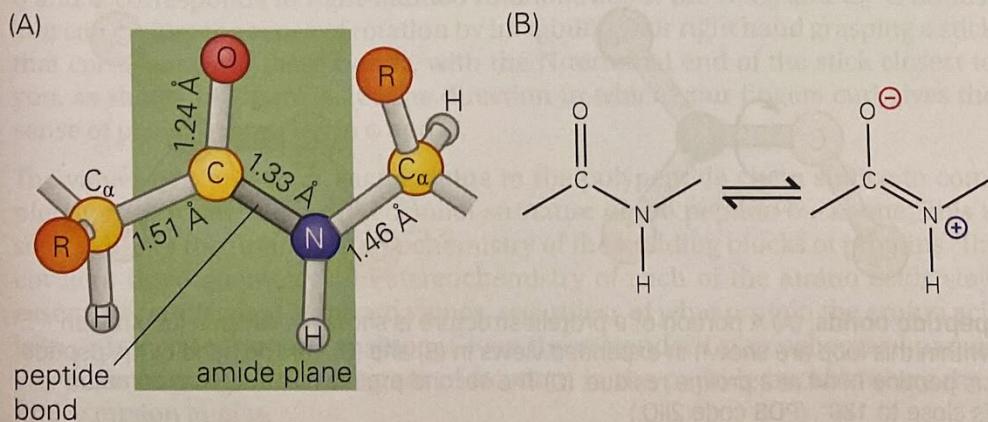
cannot be converted to D amino acids by rotations about covalent bonds. The D-form amino acids are used in certain situations, such as in the construction of bacterial cell walls, but their utilization occurs as a result of the action of specialized enzymes and not as part of the genetic decoding process.

One consequence of the chirality of amino acids is that the structural elements in proteins have a definite handedness.  $\alpha$  helices in naturally occurring proteins, for example, are right-handed. Chemists are able to synthesize artificial proteins in which all the amino acids are of the D form instead of the L form. The resulting protein structures are mirror images of their naturally occurring counterparts, with left-handed  $\alpha$  helices instead of right-handed ones.

#### 4.7 The peptide bond has partial double bond character, so rotations about it are hindered

Depending on the nature of the bond about which the rotation occurs, and the nature of the substituent groups, the energy barriers to a conformational change may be very small or insurmountably large. Rotations are strongly hindered when the covalent bond about which the rotations are being considered is a double bond, or has partial double bond character. High-resolution x-ray crystallographic studies have provided us with detailed structural information about the peptide group (Figure 4.14). Based on these studies, we know that the peptide bond (that is, the bond between the C atom in the first residue and the N atom in the second residue) and its four substituents are essentially coplanar. The four coplanar atoms of the peptide group (C, O, N, and H) form the **amide plane**, which is indicated by shading in Figure 4.14A.

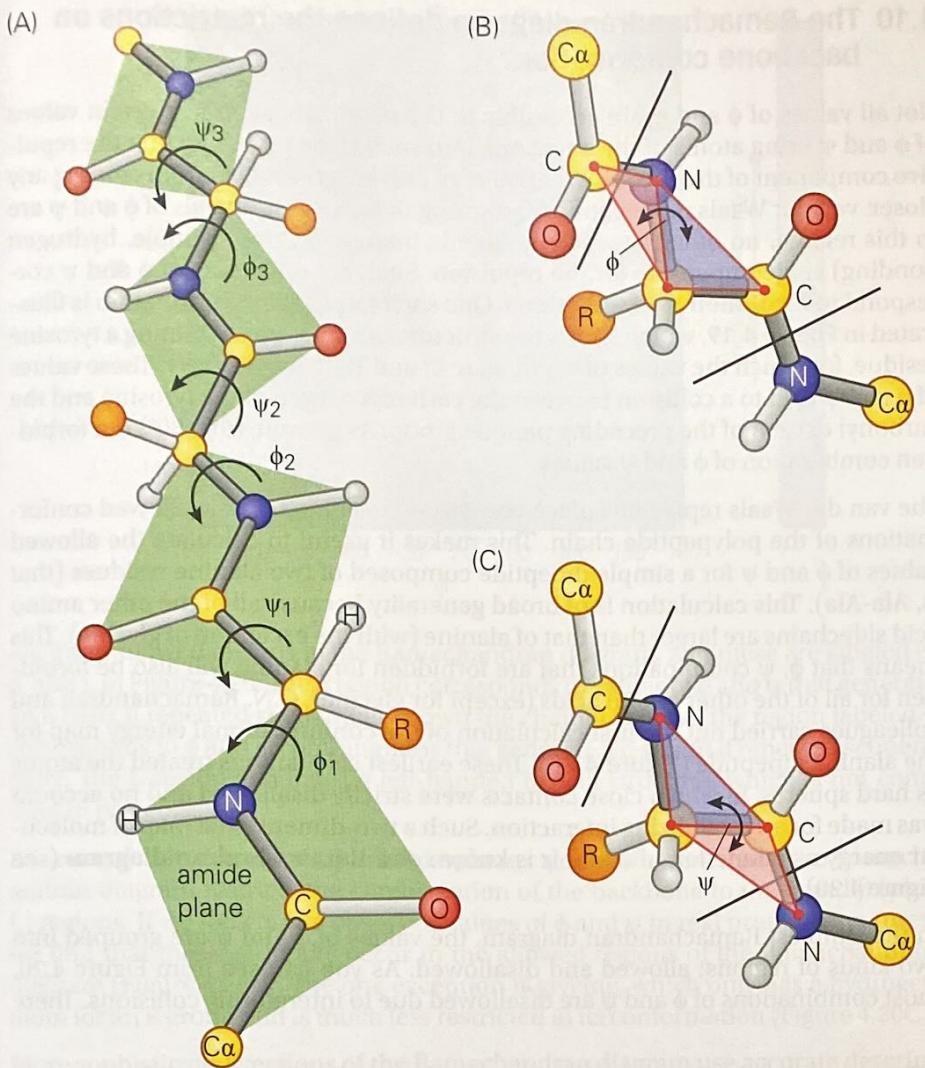
A hint regarding the origin of the planarity of the peptide group is provided by looking at Figure 4.14A, in which the lengths of the various covalent bonds in the peptide group are given. The C-N bond length ( $1.33 \text{ \AA}$ ) is shorter than the N-C $_{\alpha}$  bond length ( $1.46 \text{ \AA}$ ). This  $0.13\text{-}\text{\AA}$  difference in bond length arises because the C-N bond has partial double bond character, which can be thought of as arising from resonance between the two structures shown in Figure 4.14B. There is, as a consequence, a very large energy penalty for distorting the peptide group away from planarity.



#### Amide plane

The junction between two amino acid residues in a protein is formed by the C=O and N-H groups of the first and second residues, respectively. The four atoms in the peptide group are coplanar, and define the amide plane, which is illustrated in Figure 4.14.

**Figure 4.14 The planarity of the peptide group.** (A) The structure of the peptide group is shown, with bond lengths indicated. The amide plane, which is restricted to planarity because of the partial double bond character of the C-N bond, is shaded. (B) Resonance structures for the peptide group. The nitrogen atom in the resonance structure on the right is shown with a positive charge. In an actual peptide group, the nitrogen has partial negative charge because it withdraws electrons from the hydrogen. (A, adapted from R.E. Marsh and J. Donohue, *Adv. Protein Chem.* 22: 235–256, 1967.)

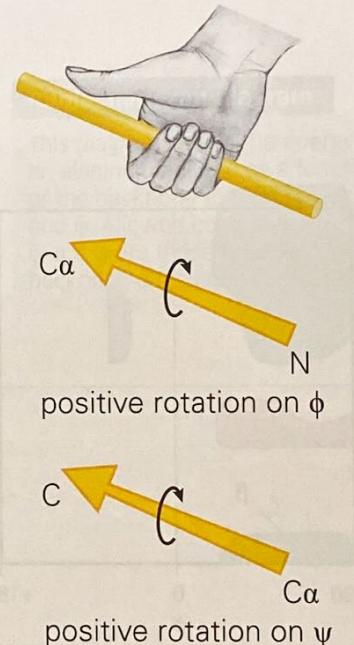


**Figure 4.17** The backbone torsion angles  $\phi$  and  $\psi$ . (A) This diagram illustrates the swivel points of the peptide backbone. (B, C) The definition of the backbone torsion angles  $\phi$  and  $\psi$ . The  $\phi$  torsion angle corresponds to rotation about the  $N-C_\alpha$  bond. The conformation shown here corresponds to a value of  $\phi = 180^\circ$  and has the  $C$  atom of the carbonyl group *trans* with respect to the  $C$  atom of the carbonyl group of the previous residue. A value of  $\phi = 0^\circ$  would have these  $C$  atoms eclipsed. The  $\psi$  torsion angle corresponds to rotation about the  $C_\alpha-C$  bond. The conformation shown here is for  $\psi = 180^\circ$  and has the backbone  $N$  atom *trans* with respect to the backbone  $N$  atom of the following residue. A value of  $\psi = 0^\circ$  would have these  $N$  atoms eclipsed. The black lines indicate the boundaries between the residues. (Adapted from J.S. Richardson, *Adv. Protein Chem.* 34: 167–339, 1981.)

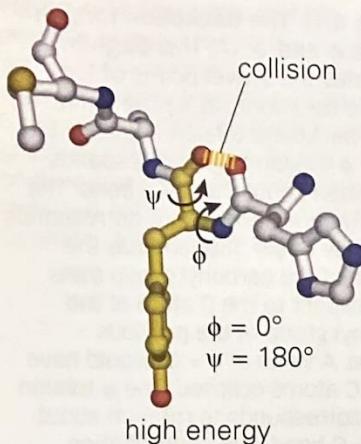
## 4.9 The backbone torsion angles $\phi$ (phi) and $\psi$ (psi) determine the conformation of the protein chain

Because of the planarity of the peptide groups, a protein backbone only has freedom to rotate at the points where the peptide groups meet (Figure 4.17A). These swivel points are at each  $C_\alpha$  atom of the peptide chain. The  $C_\alpha$  atom is connected by single bonds to the amide nitrogen ( $N$ ) and the carbonyl carbon ( $C$ ) of the same residue. Rotation about the  $N-C_\alpha$  and  $C_\alpha-C$  bonds define two torsion angles, denoted  $\phi$  and  $\psi$ , respectively (Figure 4.17B). By convention, positive rotation in  $\phi$  and  $\psi$  corresponds to right-handed rotations about the  $N-C_\alpha$  and  $C_\alpha-C$  bonds. You can picture the sense of rotation by imagining your right hand grasping a stick that corresponds to these bonds, with the N-terminal end of the stick closest to you, as shown in Figure 4.18. The direction in which your fingers curl gives the sense of positive rotations in  $\phi$  and  $\psi$ .

The values of  $\phi$  and  $\psi$  for each residue in the polypeptide chain suffice to completely specify the three-dimensional structure of the peptide backbone. This is so because of the uniform stereochemistry of the building blocks of proteins. The covalent bond geometry and stereochemistry of each of the amino acids stays essentially constant at standard values, regardless of what protein the amino acid is in or in what sequence position. Given these standard stereochemical parameters, the only significant degrees of freedom in the peptide backbone are the  $\phi$  and  $\psi$  torsion angles.



**Figure 4.18** Right-hand rule for determining the sense of  $\phi$  and  $\psi$ . Imagine grasping a rod representing the protein backbone with the fingers of your right hand. The thumb of your hand is aligned with the direction of the  $N-C_\alpha$  bond (for  $\phi$ ) or the  $C_\alpha-C$  bond (for  $\psi$ ). The direction in which your fingers are curled gives the sense of positive rotation in  $\phi$  and  $\psi$ .



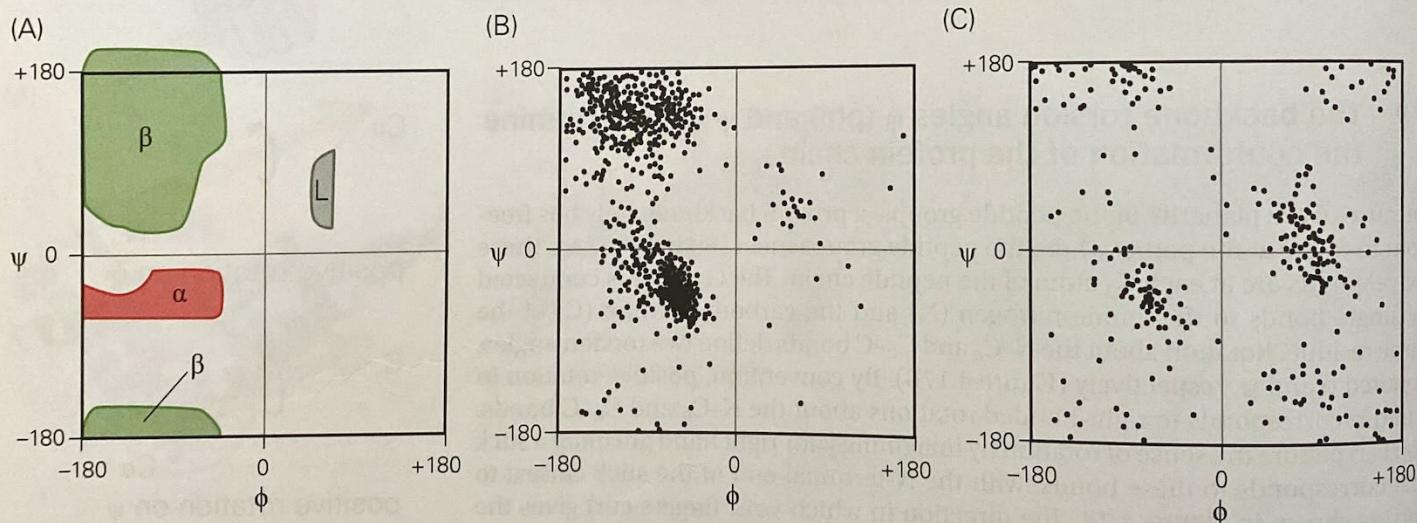
**Figure 4.19 A forbidden combination of  $\phi$  and  $\psi$ .** This particular combination of backbone torsion angles, with  $\phi = 0^\circ$  and  $\psi = 180^\circ$  for the tyrosine residue, results in a collision between the carbonyl oxygen of the tyrosine and the carbonyl oxygen of the histidine preceding it. This combination of  $\phi$  and  $\psi$  is therefore disallowed.

#### 4.10 The Ramachandran diagram defines the restrictions on backbone conformation

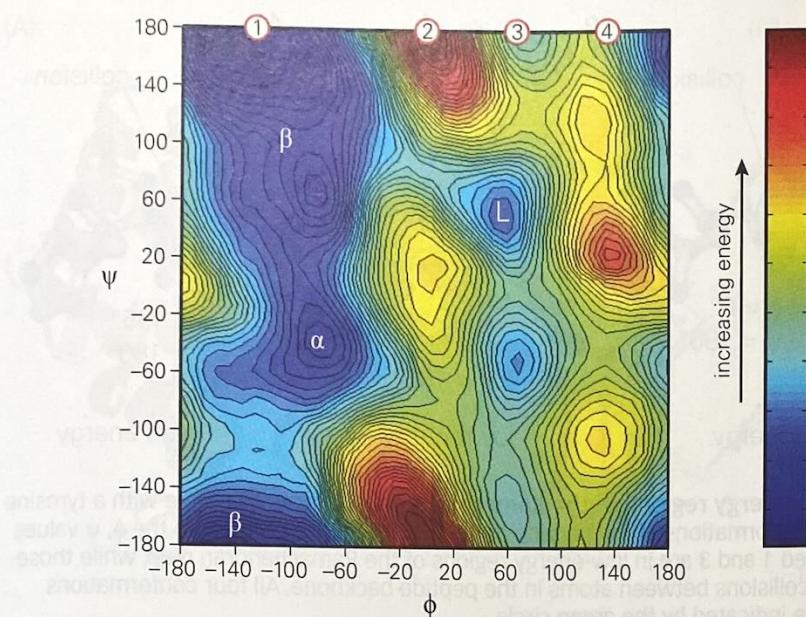
Not all values of  $\phi$  and  $\psi$  are accessible to the peptide backbone. Certain values of  $\phi$  and  $\psi$  bring atoms of the backbone into such close proximity that the repulsive component of the van der Waals energy prevents the atoms from coming any closer. van der Waals repulsions are so strong that, once the values of  $\phi$  and  $\psi$  are in this regime, no other favorable energetic interaction (for example, hydrogen bonding) can compensate for the repulsion. Such combinations of  $\phi$  and  $\psi$  correspond to **forbidden conformations**. One such forbidden conformation is illustrated in Figure 4.19, which shows the structure of a peptide containing a tyrosine residue, for which the values of  $\phi$  and  $\psi$  are  $0^\circ$  and  $180^\circ$ , respectively. These values of  $\phi$  and  $\psi$  lead to a collision between the carbonyl oxygen of the tyrosine and the carbonyl oxygen of the preceding peptide group. As a result,  $(0^\circ, 180^\circ)$  is a forbidden combination of  $\phi$  and  $\psi$  values.

The van der Waals repulsions place very strong restraints on the allowed conformations of the polypeptide chain. This makes it useful to calculate the allowed values of  $\phi$  and  $\psi$  for a simple dipeptide composed of two alanine residues (that is, Ala-Ala). This calculation is of broad generality because all of the other amino acid sidechains are larger than that of alanine (with the exception of glycine). This means that  $\phi, \psi$  combinations that are forbidden for alanine will also be forbidden for all of the other amino acids (except for glycine). G.N. Ramachandran and colleagues carried out the first calculation of the conformational energy map for the alanine dipeptide (Figure 4.20). These earliest calculations treated the atoms as hard spheres, in which close contacts were strictly disallowed and no account was made for any stabilizing interaction. Such a two-dimensional map of molecular energy as a function of  $\phi$  and  $\psi$  is known as a **Ramachandran diagram** (see Figure 4.20).

In the simplest Ramachandran diagram, the values of  $\phi$  and  $\psi$  are grouped into two kinds of regions: allowed and disallowed. As you can see from Figure 4.20, most combinations of  $\phi$  and  $\psi$  are disallowed due to interatomic collisions. There



**Figure 4.20 The Ramachandran diagram.** (A) The values of  $\phi$  and  $\psi$  for an alanine dipeptide are given by the two axes of the diagram. For each combination of  $\phi$  and  $\psi$ , the structure of the dipeptide is calculated. If the structure has interatomic collisions, combinations of  $\phi$  and  $\psi$  are shown in color and correspond to  $\alpha$ -helical conformations (red),  $\beta$  strands (green), and left-handed symmetry. The  $\beta$  region, therefore, appears at the top and the bottom of the diagram. (B) The actual values of  $\phi$  and  $\psi$  from the crystal structures of several proteins are plotted. Each dot represents the values of  $\phi$  and  $\psi$  from the (C) Values of  $\phi$  and  $\psi$  for glycine residues in several different proteins. Glycine residues can access conformations that are forbidden to the other amino acids. (Adapted from C. Brändén and J. Tooze, Introduction to Protein Structure, 2nd ed. New York: Garland Science, 1999, and also J.S. Richardson, *Adv. Protein Chem.* 34: 167–339, 1981.)



**Figure 4.21** A sophisticated version of the Ramachandran diagram.

Shown here is a Ramachandran diagram for the alanine dipeptide that uses a sophisticated evaluation of the energy of each conformation. Hydrogen bonding is accounted for, as is the interaction of the peptide with water molecules. Low- and high-energy regions are in blue and red, respectively. Notice that the lowest energy regions of this diagram are the same as those in Figure 4.20A. The conformations of a peptide with  $\phi$ ,  $\psi$  values corresponding to the points labeled 1–4 are shown in Figure 4.22. (Adapted from D.S. Chekmarev, T. Ishida, and R.M. Levy, *J. Phys. Chem. B* 108: 19487–19495, 2004. With permission from the American Chemical Society.)

are three allowed regions in the Ramachandran diagram, and these are labeled  $\alpha$ ,  $\beta$ , and L in Figure 4.20. The  $\alpha$  region corresponds to values of  $\phi$  and  $\psi$  that generate an  $\alpha$  helix if repeated sequentially down the chain. Likewise, the region labeled  $\beta$  corresponds to  $\phi$  and  $\psi$  combinations that generate a  $\beta$  strand if repeated sequentially. The L region generates a left-handed helical structure, which is not common in proteins.

It is remarkable that the simple hard-sphere calculation underlying the Ramachandran diagram restricts the conformation of the backbone to the  $\alpha$  and  $\beta$  (plus L) regions. If we look at the observed values of  $\phi$  and  $\psi$  in real protein structures, we find that these generally occur in the allowed regions of the Ramachandran diagram (Figure 4.20B). The one exception is glycine, which only has a hydrogen atom for an R group and is much less restricted in its conformation (Figure 4.20C).

More sophisticated versions of the Ramachandran diagram use accurate descriptions of the van der Waals energy and include hydrogen-bonding effects and the effects of water molecules (Figure 4.21). Despite the inclusion of more interactions, the essential features of the restriction of  $\phi$ ,  $\psi$  combinations remain essentially the same. Low-energy regions of the diagram correspond to regions in which the backbone atoms do not collide with each other, whereas backbone collisions increase the energy of other conformations.

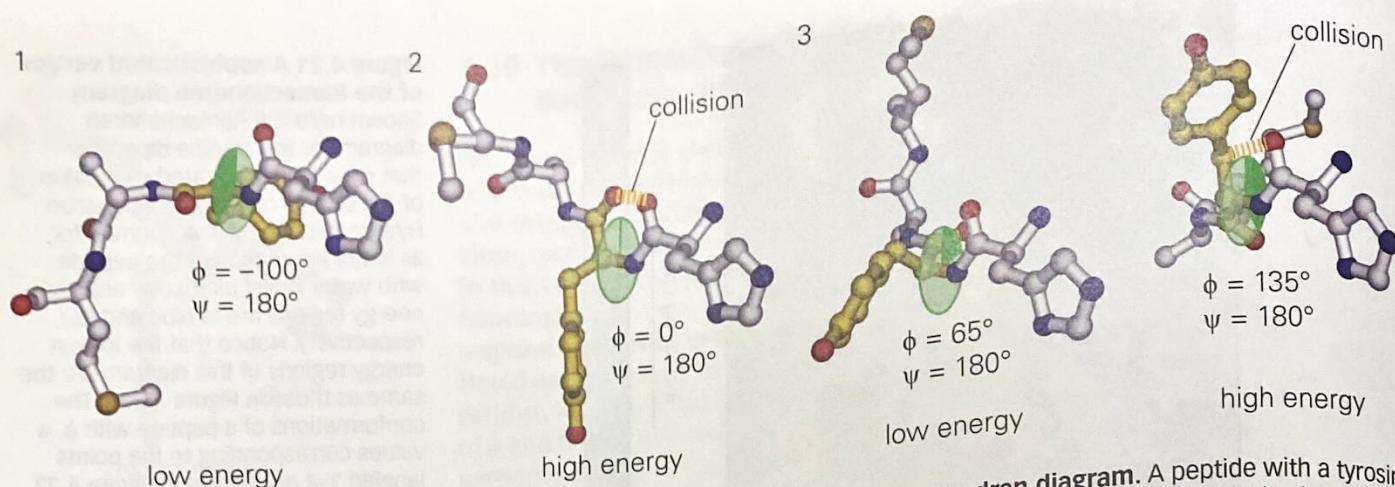
The Ramachandran maps shown in Figures 4.20 and 4.21 are calculated for an alanine dipeptide, but they help explain which conformations are disallowed for peptides containing other amino acid residues (except for glycine). This is because a conformation that is disallowed for alanine will also be disallowed for amino acids with bulkier sidechains. Four different conformations for the peptide illustrated in Figure 4.19 are shown in Figure 4.22. These conformations have different values of  $\phi$  and  $\psi$  for the central tyrosine residues. Two of the conformations have very high energy because of collisions between backbone atoms, and these correspond to disallowed regions of the Ramachandran diagram for alanine dipeptides (see Figure 4.21).

#### 4.11 $\alpha$ helices and $\beta$ strands are formed when consecutive residues adopt similar values of $\phi$ and $\psi$

The  $\alpha$  helix was first described in 1951 by Linus Pauling, who predicted that it was a structure that would be stable and energetically favorable in proteins. He made this prediction on the basis of accurate geometrical parameters that he had derived for the peptide unit from the results of crystallographic analyses of

#### Ramachandran diagram

This diagram shows the energy of an alanine dipeptide as a function of the backbone torsion angles  $\phi$  and  $\psi$ . Allowed combinations of  $\phi$  and  $\psi$  are those for which the backbone does not run into itself.



**Figure 4.22 Backbone collisions determine the high-energy regions of the Ramachandran diagram.** A peptide with a tyrosine residue (yellow) is shown here (see Figure 4.19). Four conformations of the peptide are illustrated, corresponding to the  $\phi$ ,  $\psi$  values labeled 1–4 in Figure 4.21. The  $\phi$ ,  $\psi$  combinations labeled 1 and 3 are in low-energy regions of the Ramachandran map, while those labeled 2 and 4 are in high-energy regions because of collisions between atoms in the peptide backbone. All four conformations have the same value of  $\psi$ . Changes in the value of  $\phi$  are indicated by the green circle. Changes in the value of  $\psi$  are indicated by the green circle.

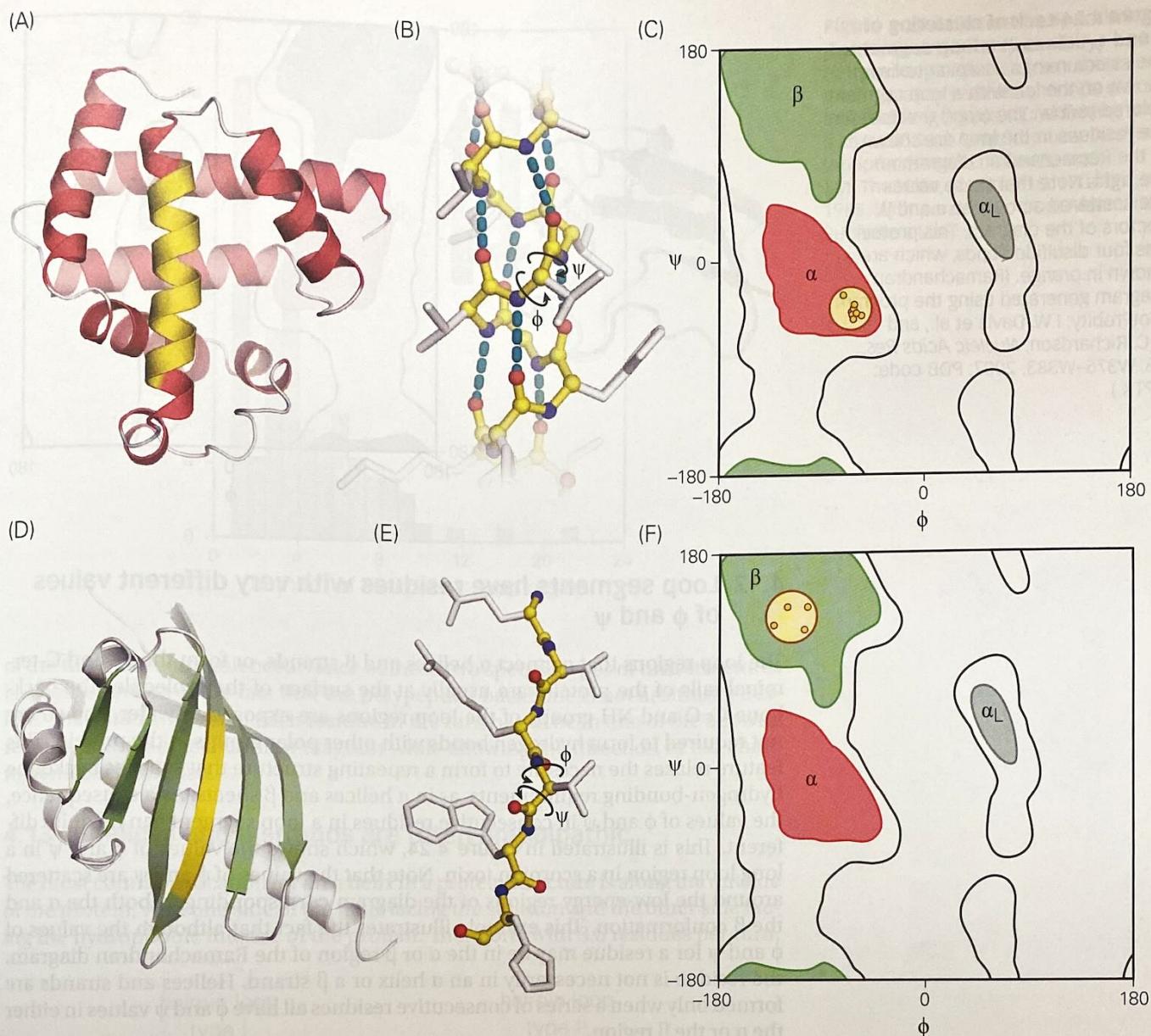
the structures of a range of small molecules. This prediction almost immediately received strong experimental support from x-ray data obtained by Max Perutz in Cambridge, UK, from hemoglobin crystals and keratin fibers. It was completely verified by John Kendrew's high-resolution structure of myoglobin, where all secondary structure is helical (see Figure 4.1).

The protein backbone forms an  $\alpha$  helix when a stretch of consecutive residues all have similar values of  $\phi$  and  $\psi$ , around  $-60^\circ$  and  $-40^\circ$ , respectively, corresponding to an allowed region in the bottom left quadrant of the Ramachandran diagram (Figure 4.23). Recall from Chapter 1 that an ideal  $\alpha$  helix has 3.6 residues per turn, with hydrogen bonds between the C=O group of one residue (denoted  $n$ ) and the NH of the residue that is four positions ahead in the chain (denoted  $n + 4$ ) (Figure 1.36). Not all  $\alpha$  helices in proteins obey this ideal rule. Note that the allowed region in the Ramachandran diagram is considerably larger than occupied by the  $\phi$  and  $\psi$  values for the near-ideal  $\alpha$  helix depicted in Figure 4.23. There is, as a consequence, some variation in the precise conformations of  $\alpha$  helices.

Variations on the  $\alpha$  helix in which the chain is either more loosely or more tightly coiled, with hydrogen bonds to residues  $n + 5$  or  $n + 3$  instead of  $n + 4$ , are called the  $\pi$  helix and  $3_{10}$  helix, respectively. The  $3_{10}$  helix has three residues per turn and contains 10 atoms between the hydrogen bond donor and acceptor, hence its name. Both the  $\pi$  helix and the  $3_{10}$  helix occur rarely and usually only at the ends of  $\alpha$  helices or as single-turn helices. They are not energetically favorable, because the backbone atoms are too tightly packed in the  $3_{10}$  helix and so loosely packed in the  $\pi$  helix that there is a hole through the middle. Only in the  $\alpha$  helix are the backbone atoms properly packed to provide a stable structure.

An  $\alpha$  helix can in theory be either right-handed or left-handed, depending on the screw direction of the chain. A left-handed  $\alpha$  helix is not, however, a favorable conformation for L amino acids due to the close approach of the sidechains and the C=O group. Thus, the  $\alpha$  helix that is observed in proteins is almost always right-handed. Short regions of left-handed  $\alpha$  helices (3–5 residues) do occur in protein structures, but only occasionally.

As for the  $\alpha$  helix,  $\beta$  strands are formed when a series of consecutive residues have similar values of  $\phi$  and  $\psi$ , except that in this case these are clustered around  $-120^\circ$  and  $+120^\circ$ , respectively (see Figure 4.23). Recall from Chapter 1 that  $\beta$  strands associate to form either parallel, antiparallel, or mixed  $\beta$  sheets, depending on

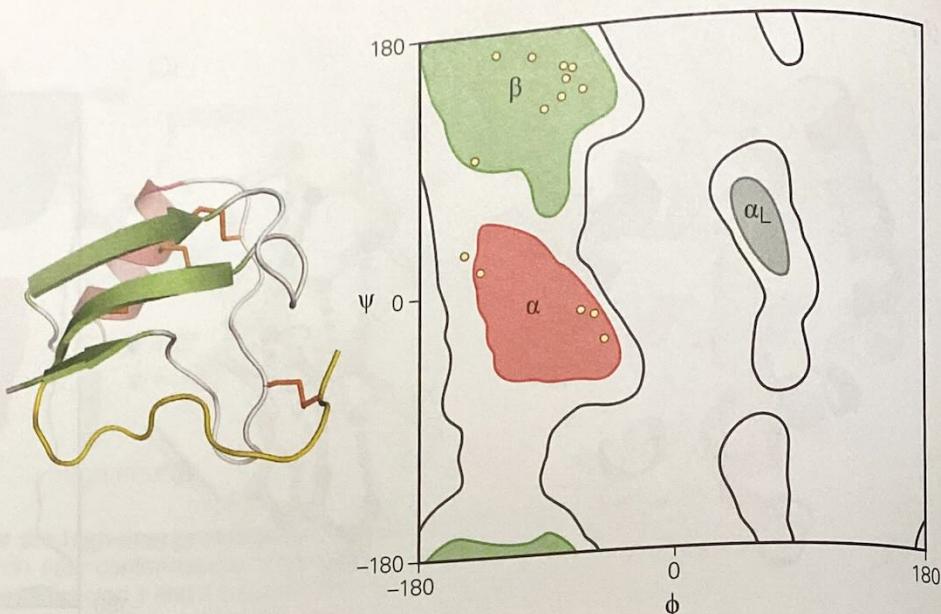


**Figure 4.23 Ramachandran diagrams for  $\alpha$  helices and  $\beta$  strands.** (A–C) An  $\alpha$  helix is formed when values of  $\phi$  and  $\psi$  close to  $-60^\circ$  and  $-50^\circ$ , respectively, are repeated along the backbone. (A) The structure of myoglobin is shown, with one helix highlighted in yellow. (B) An expanded view of the structure of this helix, with  $\phi$  and  $\psi$  indicated for a valine residue in the middle of the helix. (C) Values of  $\phi$  and  $\psi$  for all the residues in the helix are plotted on a Ramachandran diagram. Note that these values all cluster close together in the  $\alpha$  region of the diagram (yellow circle). (D–F) A  $\beta$  strand is formed when consecutive residues in the polypeptide chain have values of  $\phi$  and  $\psi$  close to  $-120^\circ$  and  $+120^\circ$ , respectively. (D) The structure of a protein containing an antiparallel  $\beta$  sheet is shown, with one strand highlighted in yellow. (E) An expanded view of the structure of this strand, with  $\phi$  and  $\psi$  indicated for a valine residue in the middle of the strand. (F) Values of  $\phi$  and  $\psi$  for four residues in the strand are plotted on a Ramachandran diagram. These values are clustered in the  $\beta$  region of the diagram (yellow circle). (Ramachandran diagrams were generated using the program MolProbity: I.W. Davis et al., and D.C. Richardson, *Nuc. Acids Res.* 35: W375–W383, 2007; PDB codes: A and B, 1A6M; D and E, 1HKX.)

whether the strands all run in the same direction or not. The backbone conformations of parallel and antiparallel  $\beta$  sheets are slightly different, but in both cases the values of  $\phi$  and  $\psi$  are clustered in the upper left quadrant of the Ramachandran diagram. The values of  $\phi$  and  $\psi$  for a strand in an antiparallel  $\beta$  sheet are illustrated in Figure 4.23.

**Figure 4.24 Lack of clustering of  $\phi$  and  $\psi$  values in a loop segment.**

The structure of a scorpion toxin is shown on the left with a loop segment colored yellow. The  $\phi$  and  $\psi$  values for the residues in the loop are shown in the Ramachandran diagram on the right. Note that these values are scattered across the  $\alpha$  and  $\beta$  sectors of the diagram. This protein has four disulfide bonds, which are shown in orange. (Ramachandran diagram generated using the program MolProbity; I.W. Davis et al., and D.C. Richardson, *Nucleic Acids Res.* 35: W375–W383, 2007; PDB code: 1PTX.)

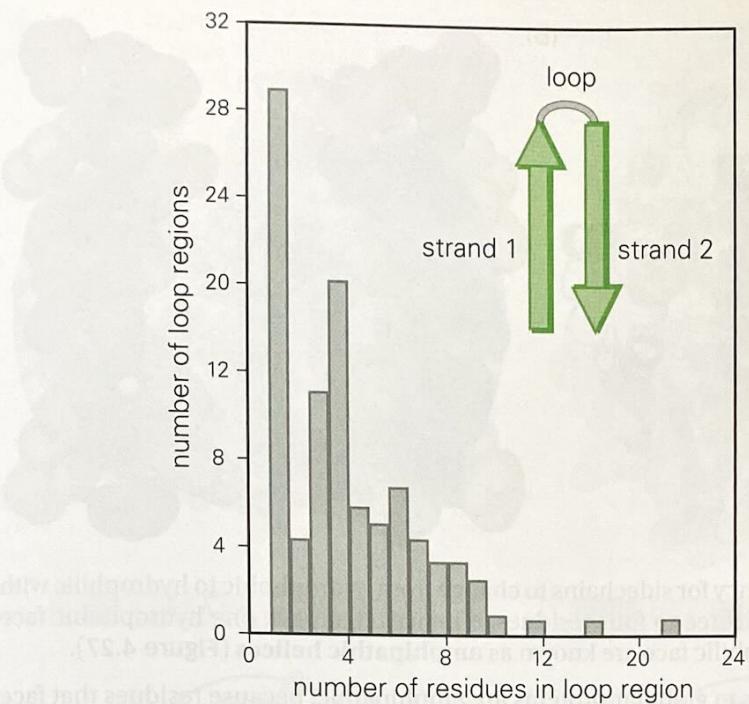


#### 4.12 Loop segments have residues with very different values of $\phi$ and $\psi$

The loop regions that connect  $\alpha$  helices and  $\beta$  strands, or form the N- and C-terminal tails of the protein, are usually at the surface of the molecule. The backbone C=O and NH groups of the loop regions are exposed to water, and so are not required to form hydrogen bonds with other polar groups in the protein. This feature relaxes the necessity to form a repeating structure that satisfies backbone hydrogen-bonding requirements, as in  $\alpha$  helices and  $\beta$  sheets. As a consequence, the values of  $\phi$  and  $\psi$  in consecutive residues in a loop segment can be quite different. This is illustrated in Figure 4.24, which shows the values of  $\phi$  and  $\psi$  in a long loop region in a scorpion toxin. Note that the values of  $\phi$  and  $\psi$  are scattered around the low-energy regions of the diagram corresponding to both the  $\alpha$  and the  $\beta$  conformation. This example illustrates the fact that although the values of  $\phi$  and  $\psi$  for a residue may lie in the  $\alpha$  or  $\beta$  region of the Ramachandran diagram, the residue is not necessarily in an  $\alpha$  helix or a  $\beta$  strand. Helices and strands are formed only when a series of consecutive residues all have  $\phi$  and  $\psi$  values in either the  $\alpha$  or the  $\beta$  region.

The scorpion toxin structure illustrated in Figure 4.24 has unusually long loop regions, with the loop highlighted in yellow being 14 residues long. We have emphasized that protein structures are stabilized by the packing of hydrophobic sidechains at the interfaces between helices and strands, which raises the question as to how the very long loops in the toxin structure are stabilized. The answer is that, in addition to the hydrophobic core, this particular protein fold is held together by four disulfide bonds that provide covalent linkages between different parts of the protein backbone (see Figure 4.24). Disulfide bonds are commonly found in secreted proteins. Disulfide bonds are unstable inside the cell because of the presence of molecules known as reducing agents, which break disulfide bonds.

Although they are irregular in conformation, loop structures in proteins quite often have specific conformations that are found in many different proteins. We shall not discuss the variety of loop conformations in detail, but simply illustrate one kind of loop—namely, that connecting adjacent antiparallel  $\beta$  strands. These are known as  **$\beta$  hairpin loops**, or **reverse turns**. As shown in Figure 4.25, such loops are usually quite short, and typically contain only four to six residues. Figure 4.26 shows two of the most frequently occurring turns—the type I turn and the type II turn—which are distinguished by the  $\phi$ ,  $\psi$  values of the two central residues

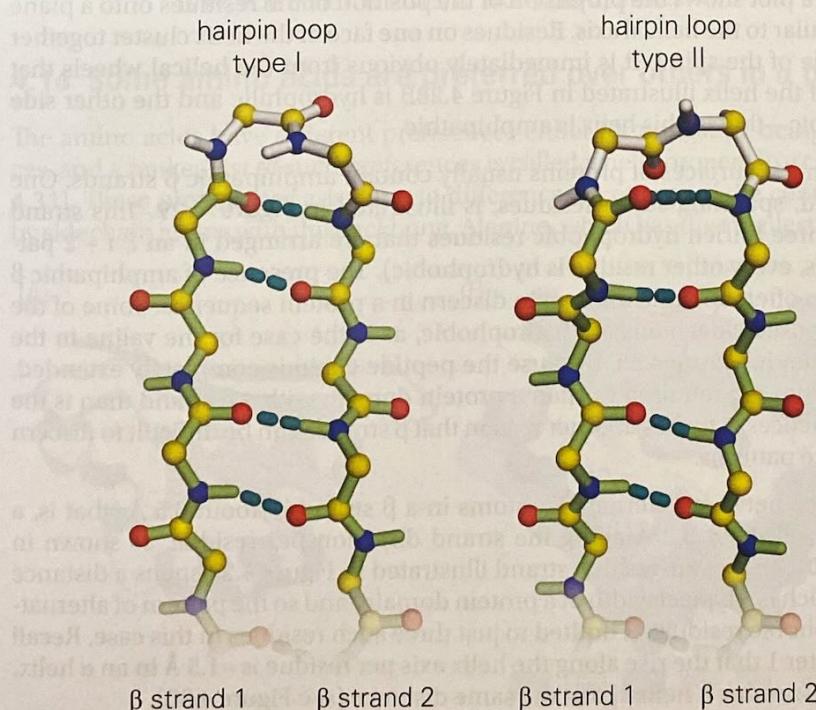


**Figure 4.25 Adjacent antiparallel  $\beta$  strands are often joined by hairpin loops.** This diagram shows the frequency with which loops of different lengths are found linking two  $\beta$  strands. Such loops are usually short. (Adapted from B.L. Sibanda and J.M. Thornton, *Nature* 316: 170–174, 1985. With permission from Macmillan Publishers Ltd.)

of the turn. The common occurrence of these two specific types of turn structures is a consequence of the fact that the polypeptide backbone is not arbitrarily flexible. The conformation of each residue in the turn has to be in one of the allowed regions of the Ramachandran diagram while allowing the direction of the chain to bend back on itself.

#### 4.13 $\alpha$ helices and $\beta$ strands are often amphipathic

The most common location for an  $\alpha$  helix in a protein structure is along the outside of the protein, with one side of the helix facing the solution and the other side facing the hydrophobic interior of the protein. Therefore, with 3.6 residues per turn,



**Figure 4.26 Two kinds of  $\beta$  hairpin loops.** The diagram shows the two most frequently occurring kinds of two-residue hairpin loops: the type I turn (left) and the type II turn (right). Bonds within the hairpin loop are shown in white. (Adapted from C. Brändén and J. Tooze, *Introduction to Protein Structure*, 2nd ed. New York: Garland Science, 1999.)