CHAPTER 3

Proteins

STUDY GOALS

After reading this chapter you will:

- understand that proteins are composed of amino acids, and know the general structure of an amino acid
- appreciate that the variability of the amino acid side-chains enables proteins with different biochemical properties to be constructed
- be able to discuss the key structural and chemical features of amino acids: enantiomeric pairs, ionization properties and polarity
- be aware of the range of modifications to amino acid structure that can occur after protein synthesis
- recognize the differences between the terms 'primary', 'secondary', 'tertiary' and 'quaternary' when referring to protein structure
- know the structure of the peptide group and the importance of the *psi* and *phi* bond angles in determining the conformation of a polypeptide
- be able to describe the α-helix and β-sheet secondary structures
- understand the structural features of fibrous and globular proteins, and be able to describe examples of both types
- recognize that quaternary structure involves the association of two or more polypeptides, and be able to describe the key features of examples of proteins with quaternary structure
- know how proteins fold and appreciate that the structures taken up are determined by the amino acid sequence
- be able to explain the link between the chemical variability of proteins and the range of different roles that proteins play in living organisms

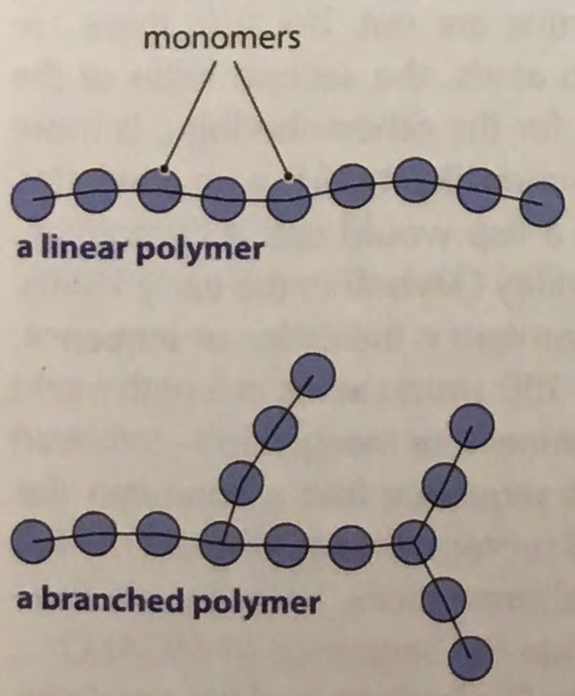


Figure 3.1 Linear and branched polymers.

In the previous chapter we examined the diversity of living organisms and the important features of prokaryotic and eukaryotic cells. Now we must look more closely at the biomolecules within those cells.

There are four types of biomolecule: proteins, nucleic acids, lipids and polysaccharides. Each has a polymeric structure, built up by linking the monomeric units together in linear or branched chains (*Fig. 3.1*). The chemical features of the monomeric units are quite different, giving each type of biomolecule its own distinctive properties. As we will see, these properties underlie the specific roles that biomolecules play in living cells. We begin with proteins.

3.1 Proteins are made of amino acids

In proteins, the monomeric units are amino acids. These are linked together to form unbranched chains called **polypeptides**. Most polypeptides are a few hundred amino acids in length, although the shortest have less than 50 amino acids (these are more correctly called **peptides**) and the longest one known, a human muscle protein called titin, has 33 445 amino acids (*Table 3.1*).

Table 3.1. Examples of human proteins

Protein	Number of amino acids	Function
Sarcolipin	31	Calcium ion transport into muscle cells
Somatotropin	51	Growth hormone
Ribonuclease A	124	Breakdown of RNA molecules
Carbonic anhydrase	130	Removal of carbon dioxide from tissues
β-globin	146	Component of hemoglobin, which carries oxygen in the bloodstream
Myoglobin	154	Utilization of oxygen by muscle tissue
Tissue plasminogen activator	527	Part of the blood clotting system
Hsp70	641	Molecular chaperone, helps other proteins adopt their correct structures
Keratin type II	644	Component of hair and cytoskeleton
Type 1 collagen	1464	Component of tendons, ligaments and bones
Dystrophin	3685	Part of the internal skeleton of muscle cells
Titin	33 445	Structural component of muscle

3.1.1 Twenty different amino acids are used to make proteins

Table 3.2. Amino acids

	Abbreviation	
Amino acid	Three-letter	One-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Polypeptides contain mixtures of 20 different amino acids (*Table 3.2*). Biochemists always use the common names for these amino acids, most of which were assigned when the individual amino acids were first discovered. Asparagine, for example, is named after asparagus, because it was first extracted from asparagus leaves, way back in 1806. Its full chemical name is 2-amino-3-carbamoylpropanoic acid. Each amino acid is also given a three-letter and one-letter abbreviation. The three-letter abbreviations are easy to remember as most are simply the first three letters of the full name, The exceptions are 'trp' for tryptophan (it is not 'try' because this might be confused with 'tyr' for tyrosine) and 'asn' and 'gln' for asparagine and glutamine ('asp' and 'glu' are used for aspartic acid and glutamic acid).

The one-letter abbreviations can be more difficult to learn. Eleven of these are the first letter of the full name (e.g. A for alanine), but nine are not, because there are not enough initial letters to go round. For two amino acids, the second letter of the name is used (R for arginine, and Y for tyrosine), but for the others the logic is more quirky. Phenylalanine is abbreviated to F because it sounds like it starts with that letter. Tryptophan is W, supposedly because a person with a lisp would call it twptophan. These abbreviations were assigned by Dr Margaret Oakley Dayhoff in the early 1960s, and have a very important purpose. A key feature of a protein is the order, or **sequence**, of amino acids in its polypeptide chain. A polypeptide 300 amino acids in length might therefore have the sequence methionine–glycine–alanine–leucine–glycine– followed by another 295 amino acids. If we wish to enter this sequence into a computer (for example, to compare it with the sequence of a related protein) then typing out the full names of the amino acids, or even the three-letter abbreviations (met–gly–ala–leu–gly–) would be very time-consuming. So we abbreviate the sequence to MGALG..., which can be typed in more quickly. This is exactly why Dayhoff devised the one-letter

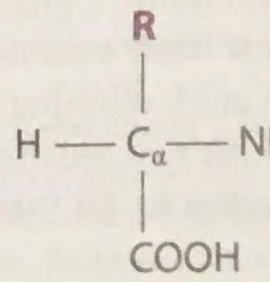


Figure 3.2 The general structure of an amino acid.
The central C is called the α-carbon.

abbreviations. She was one of the first **bioinformaticians** and the first person to use computers to study protein sequences.

Each amino acid has the same general structure (*Fig. 3.2*). This comprises a central carbon atom, called the α-carbon, to which four chemical groups are attached. These are a hydrogen atom (–H), a carboxyl group (–COOH), an amino group ($–NH_2$), and the **R group** or side-chain, which is different for each amino acid. The R groups vary greatly in complexity. For glycine, the R group is simply a hydrogen atom, whereas for phenylalanine, tryptophan and tyrosine they are large organic structures (*Fig. 3.3*). Note that proline has an unusual side-chain that includes the nitrogen of the amino group attached to the α-carbon. Because of this unusual structure, proline can introduce a kink into a polypeptide chain.

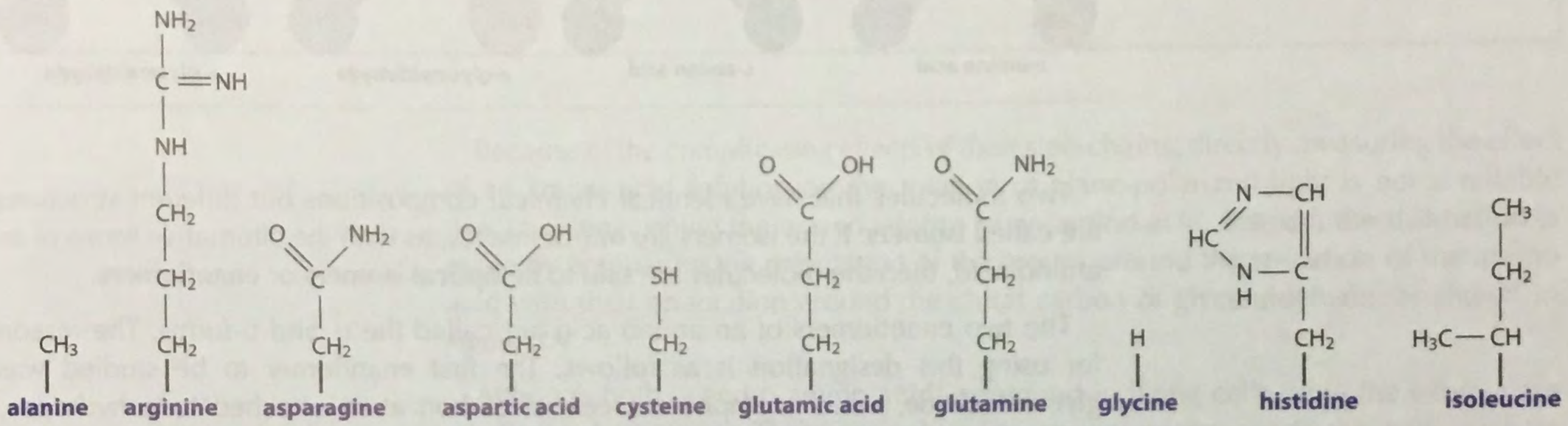


Figure 3.3 The structures of the amino acid R groups.
Note that the entire structure of proline is shown, not just its R group. This enables you to see the unusual structure of proline, in which the R group forms a bond not just to the α-carbon but also with the amino group attached to this carbon.

3.1.2 The biochemical features of amino acids

The differences between the side-chains mean that although all amino acids have the same general structure, each has its own specific chemical properties. This fact is of fundamental importance in biochemistry because it means that by combining amino acids together in different sequences, proteins with vastly different chemical features can be constructed. Later in this chapter we will examine how these different chemical features enable proteins to play a broad range of roles in living cells. First, we must understand the properties of amino acids in more detail.

There are L- and D-forms of each amino acid

The first feature of amino acids that we must consider is their precise structure. Although shown as a flat drawing in *Figure 3.2*, in reality each amino acid has a three-dimensional configuration. To understand this configuration we must consider

the way in which chemical bonds are orientated around a carbon atom. Carbon has a **valency** of four, and so can form four single bonds. These bonds have a tetrahedral arrangement. This means that there are two versions of an amino acid, differing in the positioning of the four groups around the carbon (*Fig. 3.4A*). These two versions are mirror images and so are genuinely different, and it is not possible to go from one configuration to the other simply by rotating the molecule.

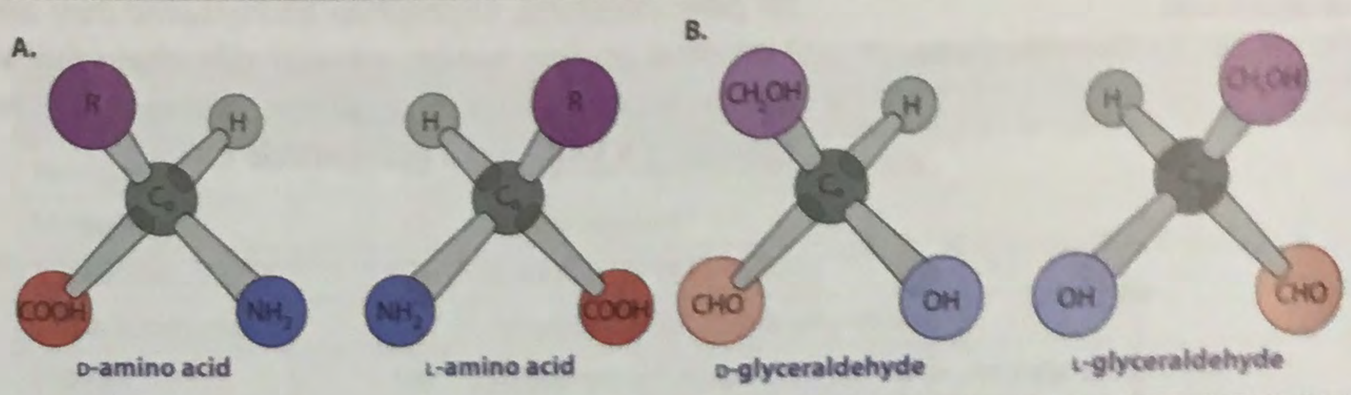


Figure 3.4 D- and L-isomers.
The D- and L-enantiomers of (A) an amino acid, and (B) glyceraldehyde, are shown. Note that each pair of enantiomers are mirror images and are not superimposable.

Two molecules that have identical chemical compositions but different structures are called **isomers**. If the isomers are mirror images, as with the alternative forms of an amino acid, then the molecules are said to be **optical isomers** or **enantiomers**.

The two enantiomers of an amino acid are called the L- and D-forms. The reason for using this designation is as follows. The first enantiomer to be studied was glyceraldehyde, which comprises a central carbon atom attached to a hydrogen, hydroxyl (–OH), formyl (–CHO) and hydroxymethyl ($–CH_2OH$) group (*Fig. 3.4B*). Like all pairs of enantiomers, the two versions of glutaraldehyde have identical chemical composition and special techniques have to be used to distinguish between them. One method is to shine plane-polarized light through a solution of the compound (*Fig. 3.5*). One enantiomer of glyceraldehyde rotates the plane of the light to the left, and the other rotates it to the right. The former was called laevo- or L-glyceraldehyde (*laevus* is Latin for 'left') and the latter dextro- or D-glyceraldehyde (*dexter* is Latin for 'right'). A carbon atom that is attached to four different groups, as in glyceraldehyde, is said to be **chiral**, from the Greek for 'hand'.

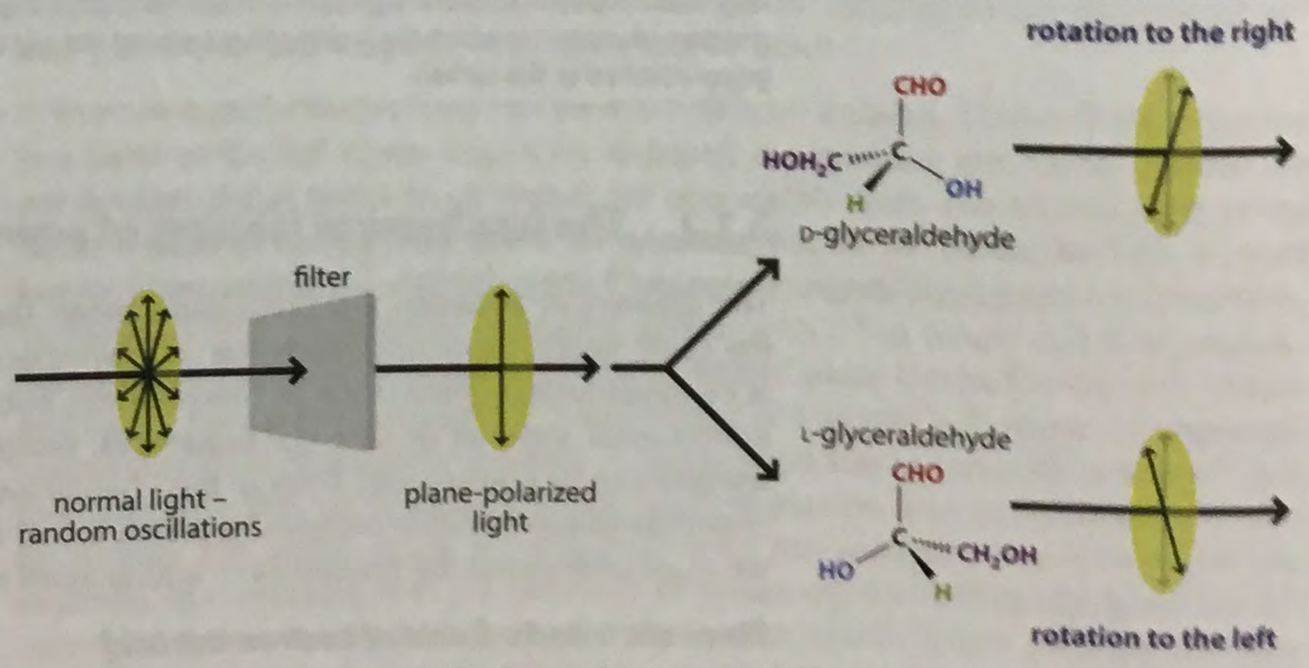


Figure 3.5 Distinguishing between the D- and L-forms of glyceraldehyde.
In normal light the waves oscillate randomly in all directions. Passing light through a special type of filter leaves only those oscillations in a single plane. When this plane-polarized light passes through a solution of an enantiomer the plane is rotated either to the right, as shown in the upper part of the diagram, or to the left, as shown below. D-glyceraldehyde causes rotation to the right, and L-glyceraldehyde causes rotation to the left.

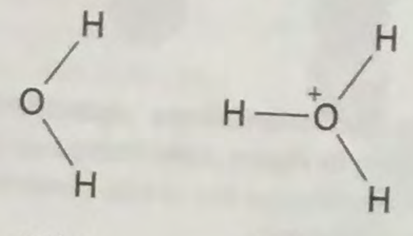
Box 3.2 The ionization of water and the pH scale

PRINCIPLES OF CHEMISTRY

Water is one of the molecules that can ionize. The chemical reaction can be described as:

$$H_2O \rightarrow H^+ + OH^-$$

In reality the H^+ ion, which is a proton, immediately combines with a second water molecule, to give a **hydronium ion** H_3O^+:



water molecule

hydronium ion

In pure water at 25°C about two in every 10^9 molecules are ionized. This corresponds to a hydronium ion concentration of 10^{-7} M.

Acids are compounds which release additional H^+ ions into a water solution. An example is hydrochloric acid, which ionizes to give a proton and a chloride ion:

$$HCl \rightarrow H^+ + Cl^-$$

Acids therefore increase the hydronium ion concentration of a solution.

Bases have the opposite effect, decreasing the hydronium ion concentration of a solution. Some bases do this directly by binding hydronium ions. Ammonia is an example, the combination between ammonia (NH_3) and a hydronium ion giving an ammonium ion (NH_4^+):

$$NH_3 + H_3O^+ \rightarrow NH_4^+ + H_2O$$

Others have an indirect effect on the hydronium ion concentration. For example, sodium hydroxide releases hydroxyl ions when it ionizes:

$$NaOH \rightarrow Na^+ + OH^-$$

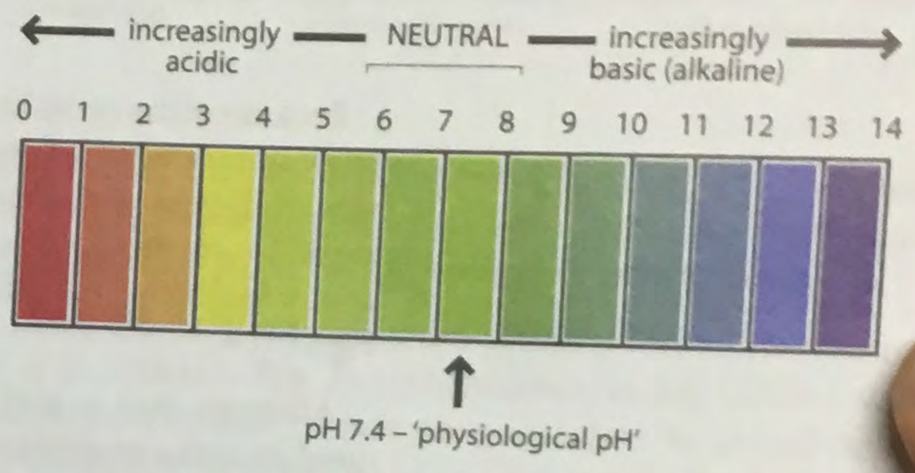
These extra hydroxyl ions combine with hydronium ions to produce non-ionized water molecules:

$$H_3O^+ + OH^- \rightarrow 2H_2O$$

The **pH** of a solution is an inverse measure of its hydronium ion concentration:

$$pH = -\log_{10}[H_3O^+]$$

where $[H_3O^+]$ means 'concentration of hydronium ions'. Pure water, with its hydronium ion concentration of 10^{-7} M, therefore has a pH of 7. An acidic solution, with a higher hydronium ion concentration than pure water, has a pH less than 7. A basic solution, with a lower hydronium ion concentration, has a pH greater than 7.



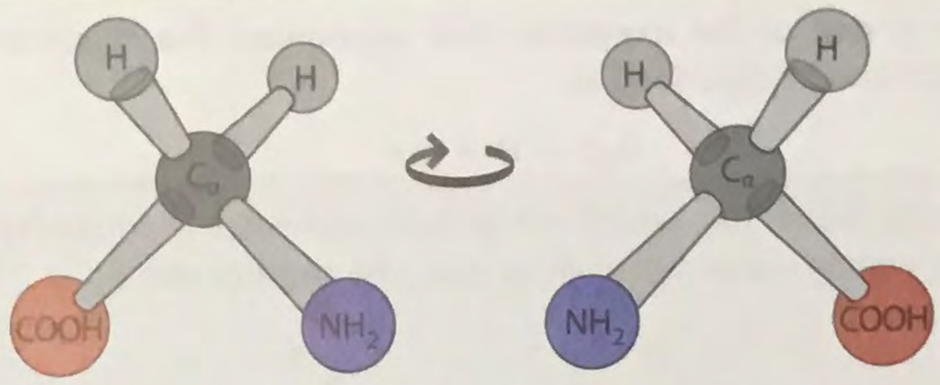
The 'physiological pH', which is the pH of most tissues in the human body, is 7.4. A slight deviation from this physiological pH can have drastically harmful effects – if the pH of human blood moves outside of the range 6.9–7.9 then the result is either coma or death. There are many reasons why the pH of living tissues is so critical, as we will see as we progress further through this book. Perhaps most importantly, changes in the pH affect the stability of some types of chemical bond, including many of the bonds that are responsible for the three-dimensional structures of biomolecules, including proteins. Changing the pH can therefore result in protein structures becoming disrupted, preventing those proteins from performing their functions in the cell.

pH 8, some of the amino groups have lost their extra protons giving molecules with a negative charge. At the pK_a for the amino group, the number of molecules with and without ionized amino groups is the same, and at pH values above the pK_a the non-ionized version of this group predominates.

Most human and plant tissues have a pH of 7.4. Which ionized form of an amino acid is present at this 'physiological pH'? For the 'typical' amino acid glycine shown in *Figure 3.7*, pH 7.4 falls well within the range at which the zwitterion predominates. This is true for all 20 of the amino acids, as indicated by the pK_a values for their carboxyl and amino groups (*Table 3.3*). These values are not all the same because they are affected by the structure of the side-chain, but they are all in the range 1.8–2.6 for the carboxyl group and 8.9–10.6 for the amino. This indicates that for all these amino acids the ionization patterns for the carboxyl and amino groups resemble those shown in *Figure 3.7*. But *Table 3.3* alerts us to a complication. Seven amino acids have side-chains that are also ionizable and which potentially could have a positive or negative charge at pH 7.4 (*Fig. 3.8*). Two of these amino acids, aspartic acid and glutamic acid,

Box 3.1 Are there two versions of every amino acid?

The definition of a chiral molecule is one that is attached to four different groups and so has a non-superimposable mirror image. In this case the carbon is said to be a chiral carbon. Not all amino acids, however, have chiral carbon atoms because there is one in which the central α-carbon is not attached to four different groups. This is the simplest of the amino acids, glycine, whose R group is a hydrogen atom. In glycine the four groups attached to the α-carbon are a carboxyl group (–COOH), an amino group ($–NH_2$), a hydrogen atom (–H), and a second hydrogen atom (–H). Glycine is therefore not chiral and does not exist as a pair of optical isomers. Glycine is just glycine and there is no such thing as D- or L-glycine.



Glycine drawn in two orientations equivalent to the D- and L-amino acids shown in *Figure 3.4A*. These two glycine molecules are not optical isomers because one can be converted to the other by rotation about the vertical axis.

Because of the complicating effects of their side-chains, directly measuring the effect of an amino acid solution on the rotation of plane-polarized light is not a reliable way of distinguishing the D- and L-forms of an amino acid. Instead, the distinction is made by comparing the orientation of the groups around the α-carbon of the amino acid with their orientation around the chiral carbon of glyceraldehyde, as shown in *Figure 3.4*.

Although both L- and D-amino acids are found in living cells, only the L-forms are used to make proteins. There are only a few special exceptions to this rule, such as some small peptides found in bacterial cell walls, which contain one or more D-amino acids.

All amino acids have ionizable groups

The second feature of amino acids that we must consider is the presence, on all amino acids, of at least two ionizable groups. In chemistry, **ionization** is the conversion of an uncharged atom or molecule into a form with an electric charge, called an **ion**. Ionization can occur by adding or removing a proton or an electron. For example, a proton has a positive charge, so adding a proton to a molecule will result in an ionic version of that molecule with a positive charge of +1. Taking away a proton will leave an ion with a negative charge of –1.

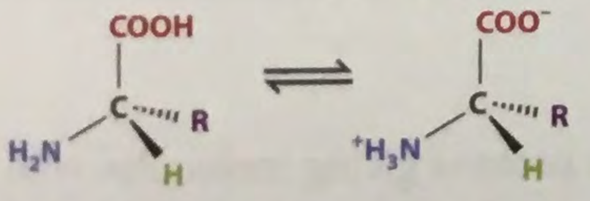


Figure 3.6 Ionization of an amino acid.

The amino acid structure shown in *Figure 3.1* has two ionizable groups. The carboxyl group (–COOH) can lose a proton and so become a negative ion ($–COO^-$), and the amino group ($–NH_2$) can gain a proton and become a positive ion ($–NH_3^+$) (*Fig. 3.6*). A molecule that has two ionized groups but no net charge is called a **zwitterion**. In chemical terms, the presence of ionizable carboxyl and amino groups on an amino acid means that these compounds can act as both weak acids and weak bases. We refer to them as being **amphoteric**.

Whether or not the carboxyl and amino groups of an amino acid are ionized depends on the pH. *Figure 3.7* illustrates this point for glycine. We see that between about pH 4 and 8 all the glycine molecules are zwitterions, with both the carboxyl and amino groups ionized. The middle of this range is called the **isoelectric point** or **pI** and at this pH the molecules carry no electrical charge. Below pH 4, some of the molecules have regained a proton, converting their $–COO^-$ groups to –COOH. These molecules therefore have a positive charge. The pH at which there are an equal number of molecules with ionized and non-ionized carboxyl groups (at this point the carboxyl groups are said to be half-dissociated) is called the pK_a of the carboxyl group. At lower pH values molecules with the non-ionized version of the carboxyl group begin to predominate. Now we move to the other end of the pH scale. Above

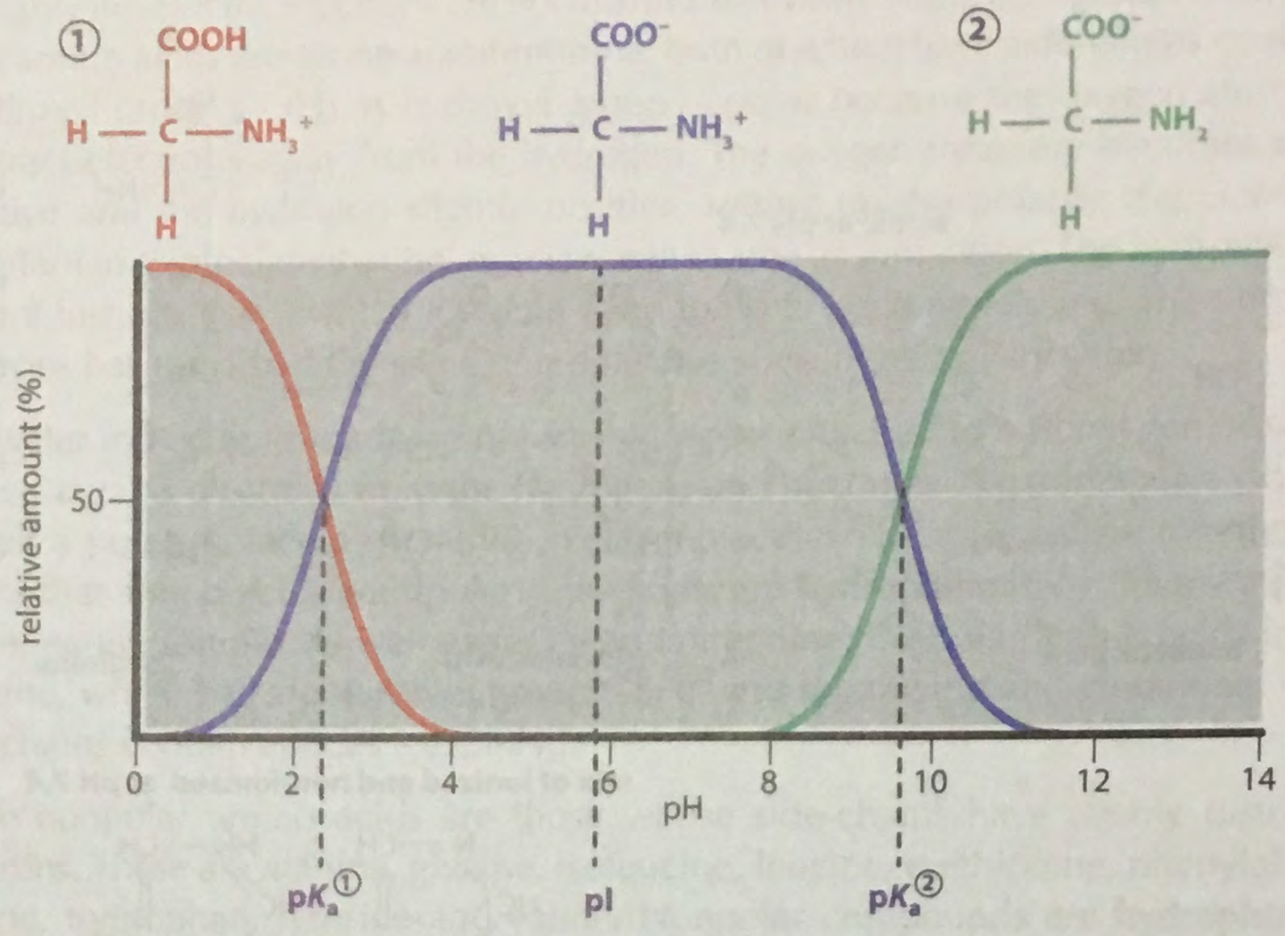


Figure 3.7 Amino acid ionization at different pH values.
The graph shows the relative amounts of the three ionized versions of glycine at pH values from 0 to 14.

have carboxyl groups in their side-chains. These carboxyls have low pK_a values (3.86 and 4.07, respectively) and are fully ionized at pH 7.4. Aspartic acid and glutamic acid therefore have acidic properties in living tissues, as their names imply. This means that they can act as acceptors of protons in biochemical reactions.

Table 3.3. Amino acid pK_a values

Amino acid	pK_a		
	Carboxyl group	Amino group	Side-chain
Alanine	2.34	9.69	
Arginine	2.01	9.04	12.48
Asparagine	2.02	8.80	
Aspartic acid	2.10	9.82	3.86
Cysteine	2.05	10.25	8.00
Glutamic acid	2.10	9.47	4.07
Glutamine	2.17	9.13	
Glycine	2.35	9.78	
Histidine	1.82	9.17	6.00
Isoleucine	2.32	9.76	
Leucine	2.33	9.74	
Lysine	2.18	8.95	10.53
Methionine	2.28	9.21	
Phenylalanine	2.58	9.24	
Proline	2.00	10.60	
Serine	2.21	9.15	
Threonine	2.09	9.10	
Tryptophan	2.38	9.39	
Tyrosine	2.20	9.11	10.07
Valine	2.29	9.72	

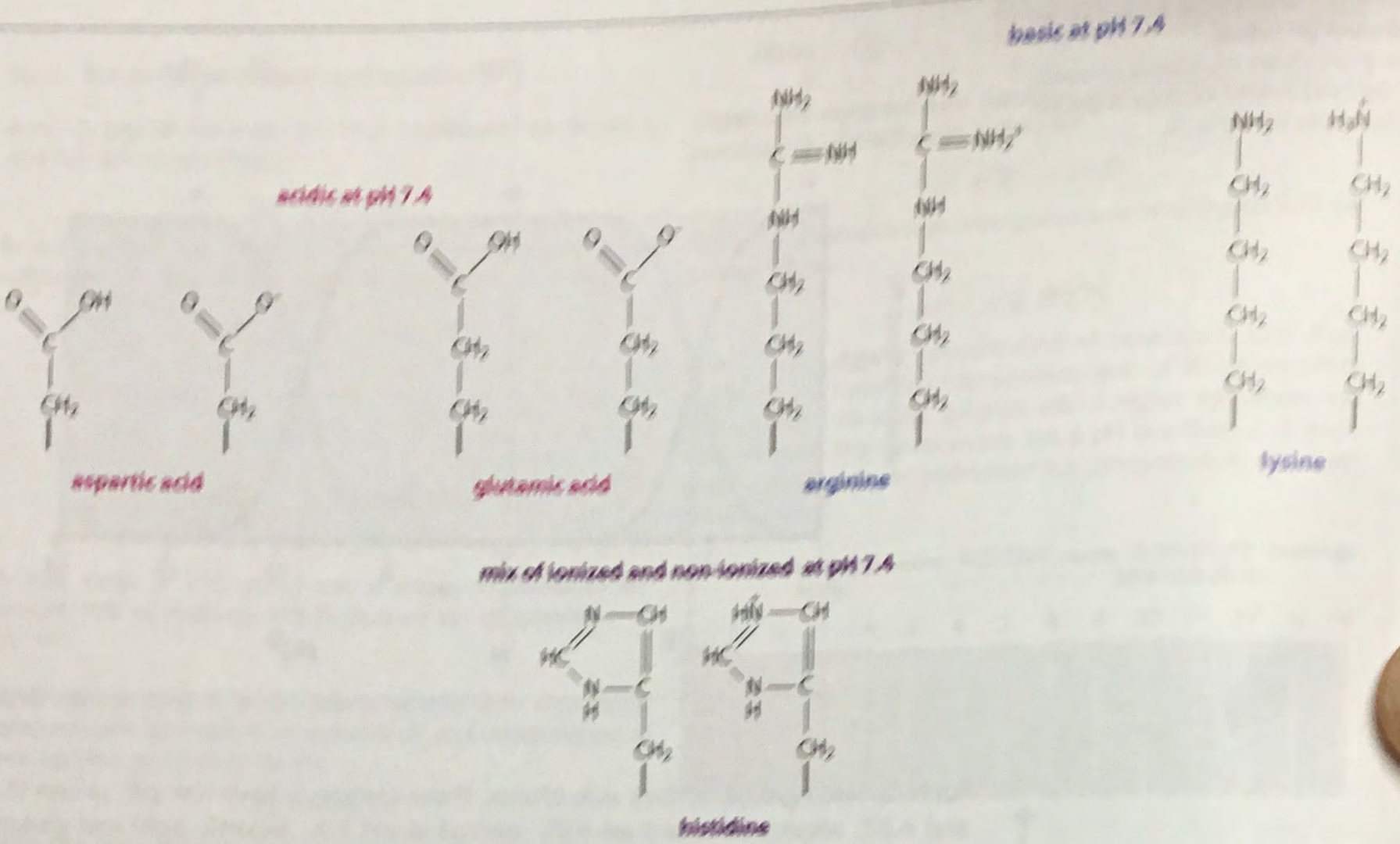


Figure 3.8 Amino acids whose side-chains are ionized at pH 7.4.

Arginine and lysine, on the other hand, both have positively charged side-chains at pH 7.4. They act as bases and can donate protons to other molecules during a biochemical reaction. The side-chains of cysteine and tyrosine also have ionizable groups, but these are largely non-ionized at pH 7.4. These two amino acids are therefore uncharged under physiological conditions. Histidine is the last in the list of amino acids with ionizable side-chains, and this one is interesting. At pH 7.4 there are significant amounts of both the ionized and non-ionized versions of the side-chain. A histidine within a protein molecule can therefore act as both a donor and acceptor of protons, a property that is exploited in a number of important biochemical reactions.

Some amino acids have polar side-chains

We have seen that some amino acids have distinctive chemical properties because their side-chains contain ionizable groups. A second distinctive feature of some R groups is their **polarity**.

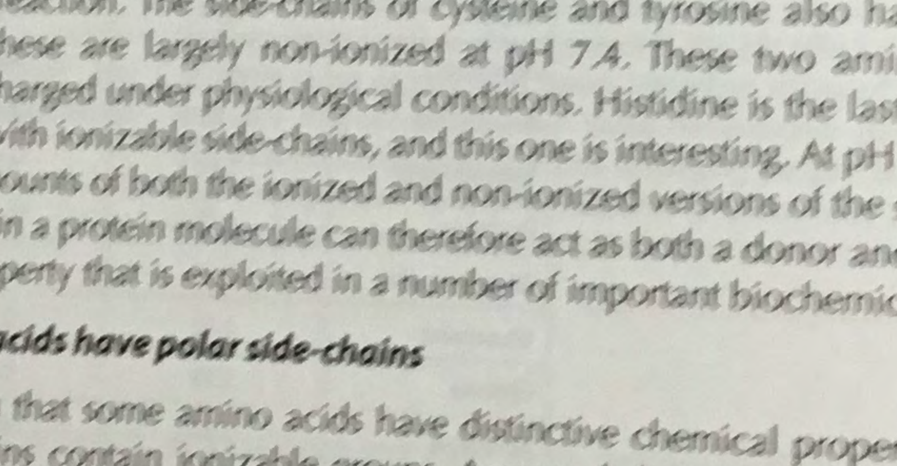


Figure 3.9 The polarity of (A) a hydroxyl group, and (B) a water molecule.
The oxygen atom tends to attract electrons away from the hydrogen(s). The oxygen therefore becomes slightly electronegative (denoted by δ^-) and the hydrogen slightly electropositive (δ^+).

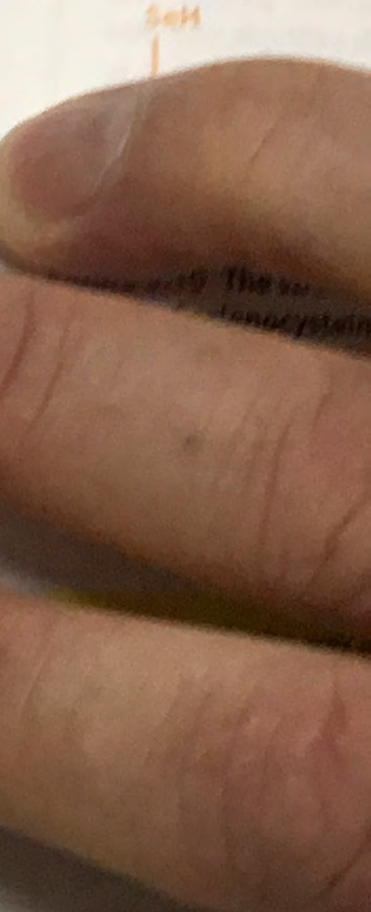


Figure 3.11 Proline and 4-hydroxyproline.

PRINCIPLES OF CHEMISTRY

Box 3.3 Types of chemical bond

Chemical bonds are inherent and essential components of all of the molecular structures that are important in biochemistry.

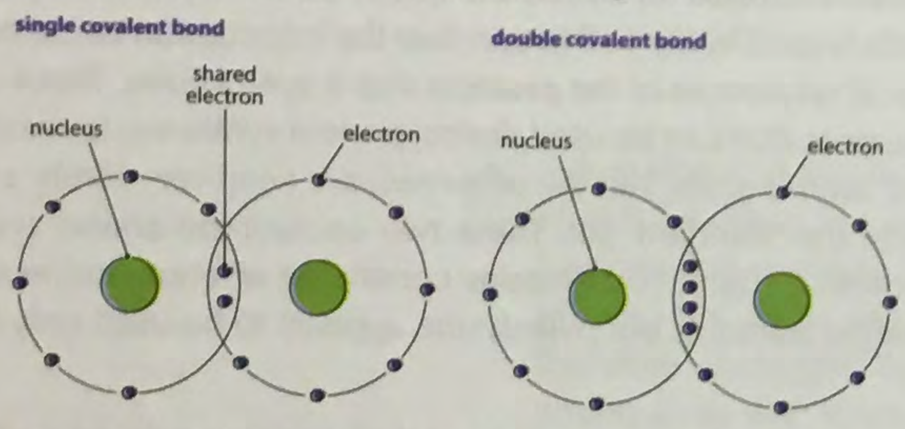
- Chemical bonds hold together the atoms in a molecule such as an amino acid or a protein.
- Chemical bonds enable interactions to form between different parts of a polymeric molecule. This might result in the polymer adopting a helical or other conformation. Similar interactions might fold the polymer into a more complex, three-dimensional structure.
- Chemical bonds enable two or more molecules to bind to one another, resulting in, for example, a multisubunit protein.

We will encounter a variety of different types of chemical bond as we study the structures of proteins and other biomolecules. The most important of these bonds are described here.

Covalent bonds

All of the bonds contained within an amino acid are covalent ones, as are the bonds in the peptide linkage. Covalent bonds are also the predominant type of bonding in nucleic acids, lipids and polysaccharides. Covalent bonds are so common in biochemistry that if the word 'bond' is used without any adjective then you can assume that the bond is a covalent one.

Covalent bonds form when two atoms share electrons. If two atoms come close enough together then two or more pairs of electrons can become shared between the two atoms. In chemical terms, the shared electrons occupy **orbitals** of both atoms. The two atoms are held tightly together, forming the bond.



If one pair of electrons is shared then a **single bond** is formed. Both of the atoms can rotate about a single bond, changing the orientation of any other bonds that those atoms have formed. A **double bond**, on the other hand, involves two pairs of shared electrons, and does not allow any rotation.

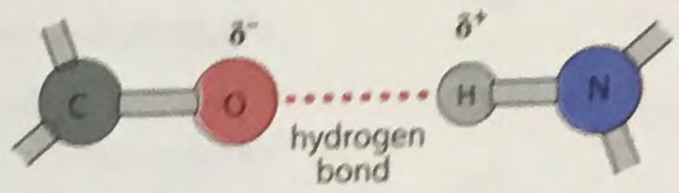
The strength of a covalent bond depends on its **bond energy**, which is a measure of the amount of energy needed to break it. The strength depends on the identities of the atoms that are linked together, and double bonds are stronger than single ones. A single bond between two carbon atoms (C–C) has a bond energy of 348 kJ mol^{-1}, whereas a carbon–carbon double bond (C=C) is 1.75× stronger, with an energy of 614 kJ mol^{-1}. A C–H bond has an energy of 413 kJ mol^{-1}, and the energy of a C–N bond is 308 kJ mol^{-1}.

Electrostatic bonds

An electrostatic bond is an interaction between positively and negatively charged chemical groups. In proteins, they form between an amino acid with a positively charged side-chain (such as lysine or arginine) and one with a negatively charged side-chain (aspartic acid or glutamic acid). They have bond energies of 6–12 kJ mol^{-1}, substantially less than covalent bonds. As well as stabilizing structures within proteins, electrostatic bonds are also important on the protein surface, where they hold together the various polypeptides of a multisubunit protein.

Hydrogen bonds

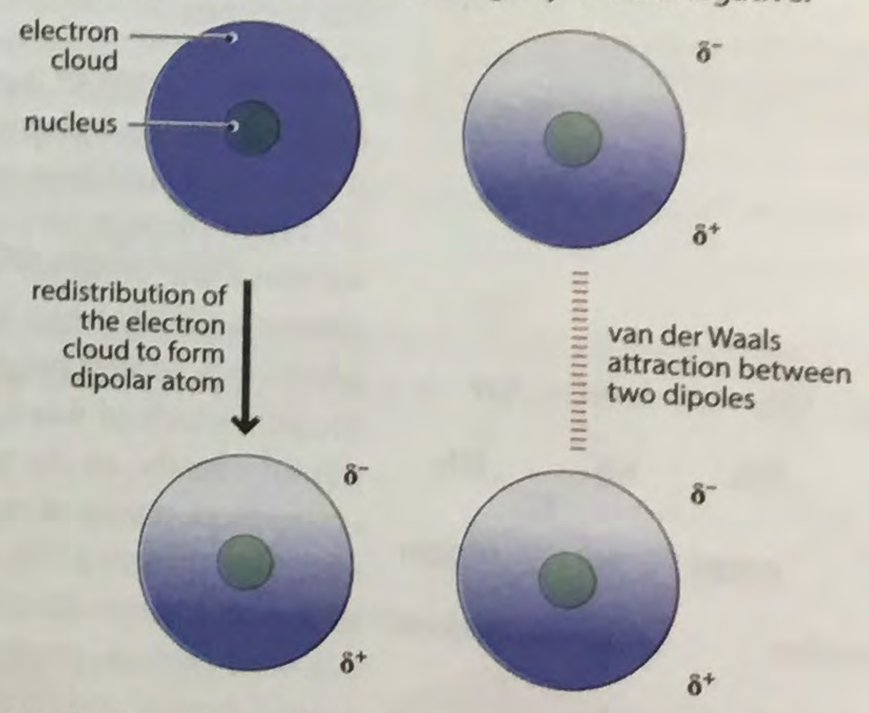
A hydrogen bond is an interaction that forms between the slightly electropositive hydrogen atom in a polar group and an electronegative atom, which might be in the same molecule or in a completely different one. The charge on the electropositive atom is designated δ^+ and on the electronegative one δ^-.



The hydrogen atom is shared between the two groups. The group to which it is more tightly linked (in this example, the -NH group) is called the 'donor' group, and the group to which it is less tightly linked (the -CO group in this example) is called the 'acceptor' group. Hydrogen bonds vary in strength depending on the atoms that are involved, but most are relatively weak. Those in biomolecules have energies between 8 and 29 kJ mol^{-1}. Often a number of hydrogen bonds participate in the same interaction between two molecules or two parts of a molecule. The resulting structure can therefore be stable at physiological temperatures, even though the individual bonds are relatively weak. Examples of such structures are the α-helix and β-sheet of proteins, and the double helix of DNA.

van der Waals forces

These are weak attractions named after the Dutch physicist Johannes van der Waals (1837–1923), who first studied them in gases and liquids. They involve temporary electrical charges that occur because of random fluctuations in the distribution of electrons around an atom. Usually the electrons are distributed evenly, in which case the atom has no electrical charge. But by chance the electron cloud can become uneven, with more electrons on one side of the atom compared with the other. This results in a **dipole**, in which one side of the atom is slightly electropositive and the other side slightly electronegative.



If two dipolar atoms are close enough together then they will attract one another, with a bond energy of about 2–4 kJ mol^{-1}.

A van der Waals attraction will last only as long as the fluctuation in the electron clouds that give rise to the dipoles is maintained. However, within a biomolecule such as a protein, there will be so many dipolar chemical groups present at any one time that there are always pairs close enough together to stabilize the biomolecular structure. The identities of the dipole pairs will constantly change, but there will always be lots of them.

3.2 The primary and secondary levels of protein structure

Proteins are traditionally looked on as having four distinct levels of structure. These levels are hierarchical, the protein being built up stage by stage, with each level of structure depending on the one below it (*Fig. 3.12*).

- The **primary structure** is the sequence of amino acids in the polypeptide.
- The **secondary structure** refers to a series of conformations, including helices, sheets and turns, that can be adopted by different parts of the polypeptide.
- The **tertiary structure** is the overall three-dimensional configuration of the protein.
- A **quaternary structure** is the association between different polypeptides to form a multisubunit protein.

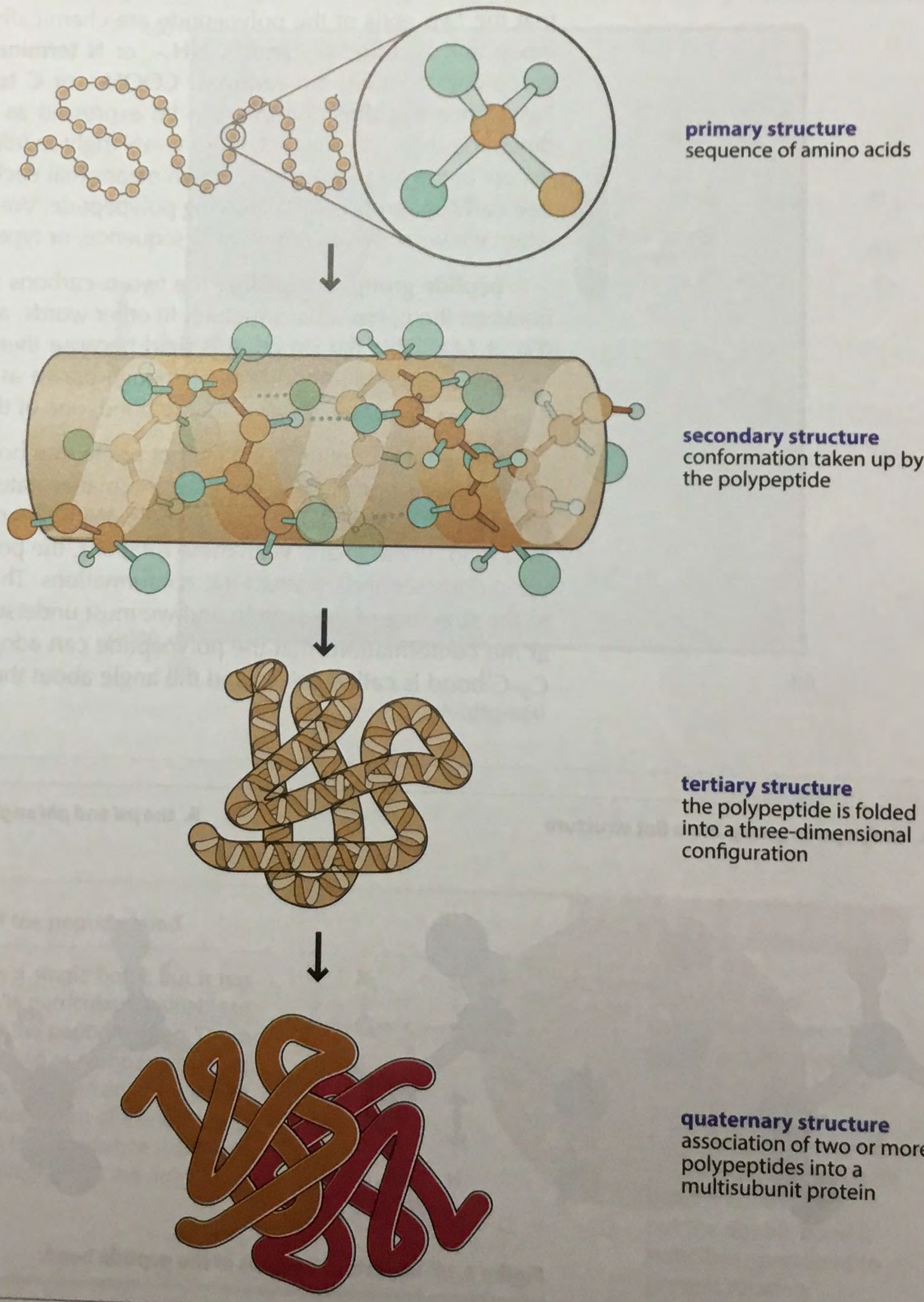


Figure 3.12 The four hierarchical levels of protein structure.

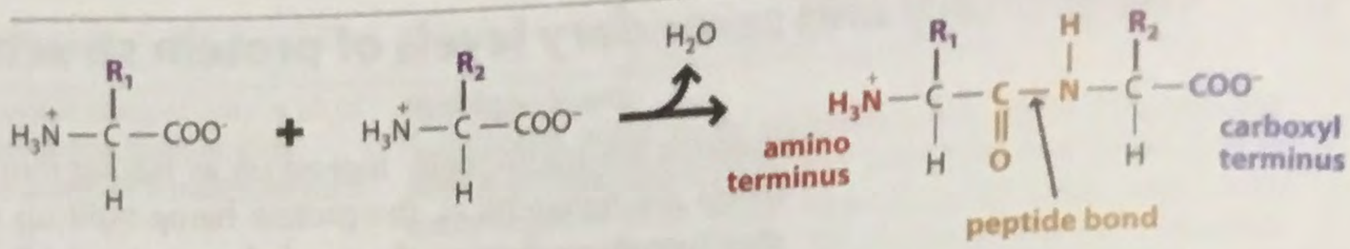


Figure 3.13 The chemical reaction that results in two amino acids becoming linked together by a peptide bond.

In this section we will study the first two of these levels of protein structure.

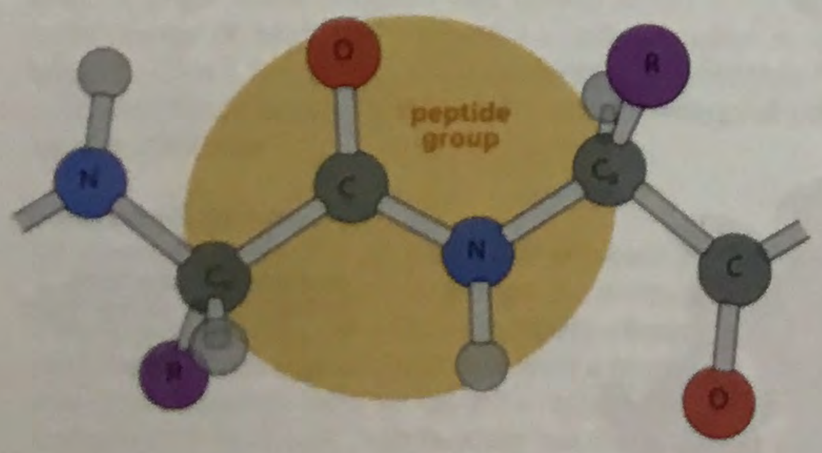
3.2.1 Polypeptides are polymers of amino acids

A polypeptide is built by linking amino acids together by **peptide bonds** (*Fig. 3.13*). Each peptide bond forms between the carboxyl and amino groups of adjacent amino acids, by a **condensation** reaction that expels a molecule of water. Note that this means that the two ends of the polypeptide are chemically distinct. One has a free amino group and is called the **amino**, **NH_2–**, or **N terminus.** The other has a free carboxyl group and is called the **carboxyl**, **COOH–**, or **C terminus**. A polypeptide therefore has a chemical direction that can be expressed as either N→C (left to right for the dipeptide shown in *Figure 3.13*) or C→N (right to left in *Figure 3.13*). Protein synthesis occurs in the N→C direction, which means that each new amino acid is added to the free carboxyl group of the growing polypeptide. We therefore use the N→C direction when we write out an amino acid sequence, or type it into a computer.

A **peptide group**, comprising the two α-carbons and the C, O, N and H atoms in between them, has a flat structure. In other words, all six atoms lie on the same plane (*Fig. 3.14A*). This flat structure is rigid because there is little opportunity for rotation around the peptide bond itself. Although drawn as a single bond, the peptide bond has some characteristics of a double bond, one of these being an inability to rotate.

Although the peptide bond cannot rotate, the bonds either side of it can. Rotation around these bonds does not alter the planar nature of the peptide group but does affect the polypeptide chain as a whole. Without any rotation, the polypeptide would be a rigid, linear chain. With these rotations, the polypeptide is able to bend and take up various secondary structural conformations. These rotations are therefore critical to the structure of the protein and we must understand them in detail before we look at the conformations that the polypeptide can adopt. The angle of rotation about the C_α–C bond is called *psi* (ψ) and the angle about the N–C_α bond is *phi* (ϕ) (*Fig. 3.14B*).

A. the peptide group has a flat structure



B. the *psi* and *phi* angles

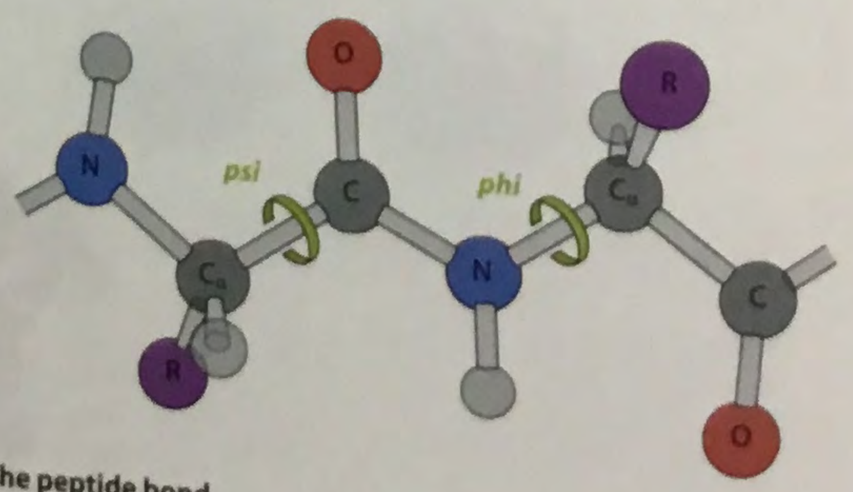


Figure 3.14 Important features of the peptide bond.

Figure 3.15
plot.
The dark blue
indicate the c
phi that are p
steric effects.
most favorabl
most combina
found in real p

Box 3.4 The
The peptid
some chara
rotate, cont
properties
involves the
in a molecul
with a doub
of the pepti

If two adjacent peptide groups are orientated in the same plane, then *psi* and *phi* are both 180°. If either bond is rotated clockwise (when looking towards the α-carbon from the other end of the bond) then the angle assigned to *psi* or *phi* increases. If the rotation is counterclockwise, then the angle decreases. The precise combination of angles for *psi* and *phi* either side of an α-carbon determines the conformation of the polypeptide at that point along its length.

It turns out that 77% of the possible combinations of *psi* and *phi* never occur, because of **steric effects**. These effects prevent two atoms from getting too close together and limit the possible conformations that any molecule can take up. The combinations of *psi* and *phi* that are allowed are shown by the **Ramachandran plot** (*Fig. 3.15*), named after G.N. Ramachandran, who led the team that first worked out the information summarized in this diagram, in 1963.

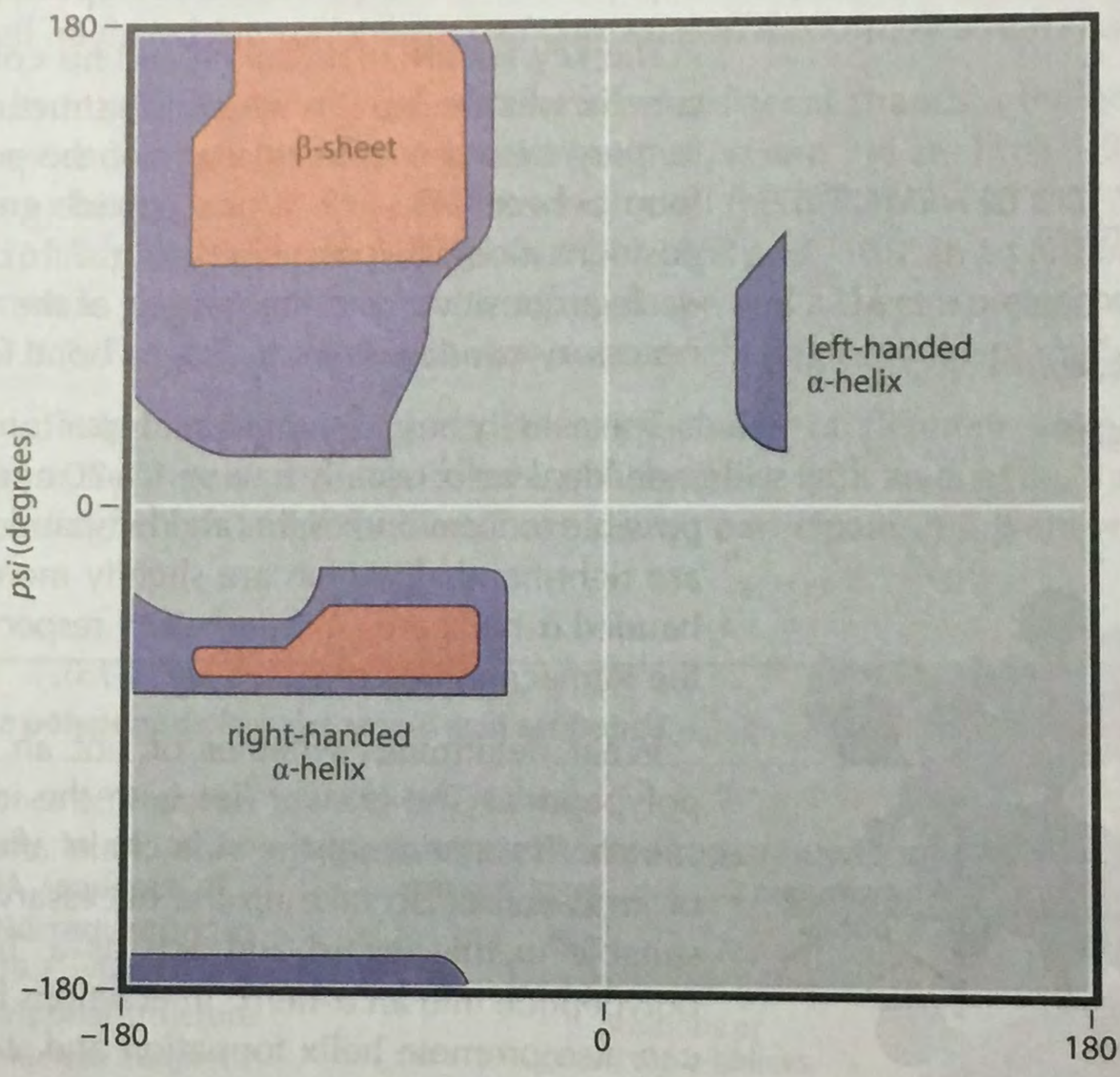
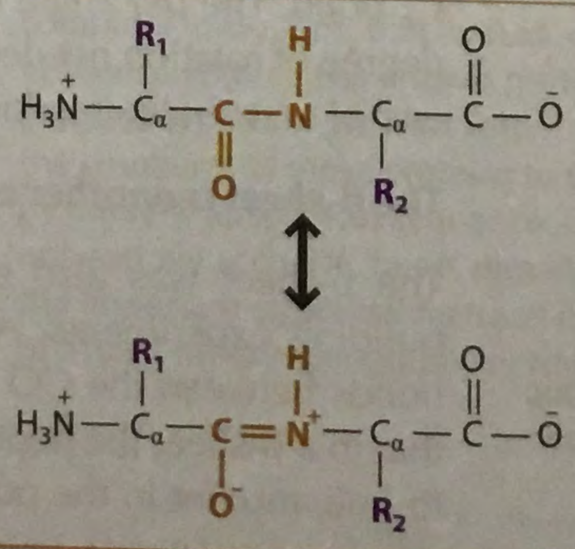


Figure 3.15 The Ramachandran plot.
The dark blue and red areas of the plot indicate the combinations of *psi* and *phi* that are possible without causing steric effects. The red areas are the most favorable regions, within which most combinations of bond angles found in real polypeptides are located. The types of secondary structure that result from the bond angles in the different regions of the plot are noted.

Box 3.4 The unusual characteristics of the peptide bond

PRINCIPLES OF CHEMISTRY

The peptide bond is usually drawn as a single bond, but it has some characteristics of a double bond. In particular, it is unable to rotate, contributing to the planarity of the peptide group. These properties arise from a process referred to as **resonance**, which involves the redistribution of electrons between adjacent atoms in a molecule. This can result in the replacement of a single bond with a double bond, and vice versa. The two resonance structures of the peptide bond are shown in the diagram on the right:



The continuous resonance between the two structures means that the peptide bond oscillates between single and double bond characteristics. The single bond predominates, but the double bond is sufficiently prevalent to prevent rotation.

3.2.2 Polypeptides can take up regular conformations

If one or both of the *psi* and *phi* angles either side of an α-carbon are different to 180°, then the polypeptide will change direction at that point. To illustrate this effect, we will study the two most common types of secondary structure found in proteins, called the α-helix and the β-sheet.

The α-helix is a common type of secondary structure

We will examine how X-ray crystallography is used to study protein structure in *Section 18.1.3*.

The α-helix was discovered by Linus Pauling, Robert Corey and Herman Branson in the late 1940s. At the time this was a rather odd type of discovery because it was not based entirely on experimental evidence. X-ray crystallography had suggested that a helical structure of some kind was a common feature of many proteins. Pauling decided to work out the structure of that helix by building models, initially just with a polypeptide chain drawn on a piece of paper. **Model-building** is still used to interpret data from X-ray crystallography, although today the models are built not with pieces of paper but with sophisticated computer programs.

The key feature that Pauling and his colleagues used to work out the details of the α-helix was the way in which the structure would be stabilized by hydrogen bonds forming between different parts of the polypeptide. In an α-helix, hydrogen bonds form between the CO of one peptide group and the NH of the peptide group four positions along the polypeptide (*Fig. 3.16*). The NH group is polar and so this hydrogen is electropositive, and the oxygen of the CO group is electronegative, providing the necessary conditions for hydrogen bond formation.

The α-helix has 3.6 amino acids per turn with the side-chains pointing outwards, an individual helix usually having 10–20 amino acids but sometimes as many as 40. It is possible to form both right- and left-handed α-helices, but almost all those in proteins are right-handed, which are slightly more stable. The *psi* and *phi* angles for a right-handed α-helix are –47° and –57°, respectively, within one of the favorable regions of the Ramachandran plot (see *Fig. 3.15*).

What determines whether or not an α-helix forms in a particular region of a polypeptide? The answer lies with the identities of the amino acids present in that stretch. The nature of the side-chain affects the ability of the bonds on either side of an α-carbon to take up the necessary *psi* and *phi* angles. Alanine is particularly suitable in this regard and acts as a 'helix former', promoting the folding of the polypeptide into an α-helix. Interactions between side-chains of different amino acids can also promote helix formation and stabilize a helix once it has formed. For such an interaction to occur, two side-chains have to be on the same face of the helix and therefore they must be 3–4 amino acids apart in the polypeptide. Electrostatic bonds between positively and negatively charged side-chains of amino acids spaced apart in this way often stabilize an α-helix.

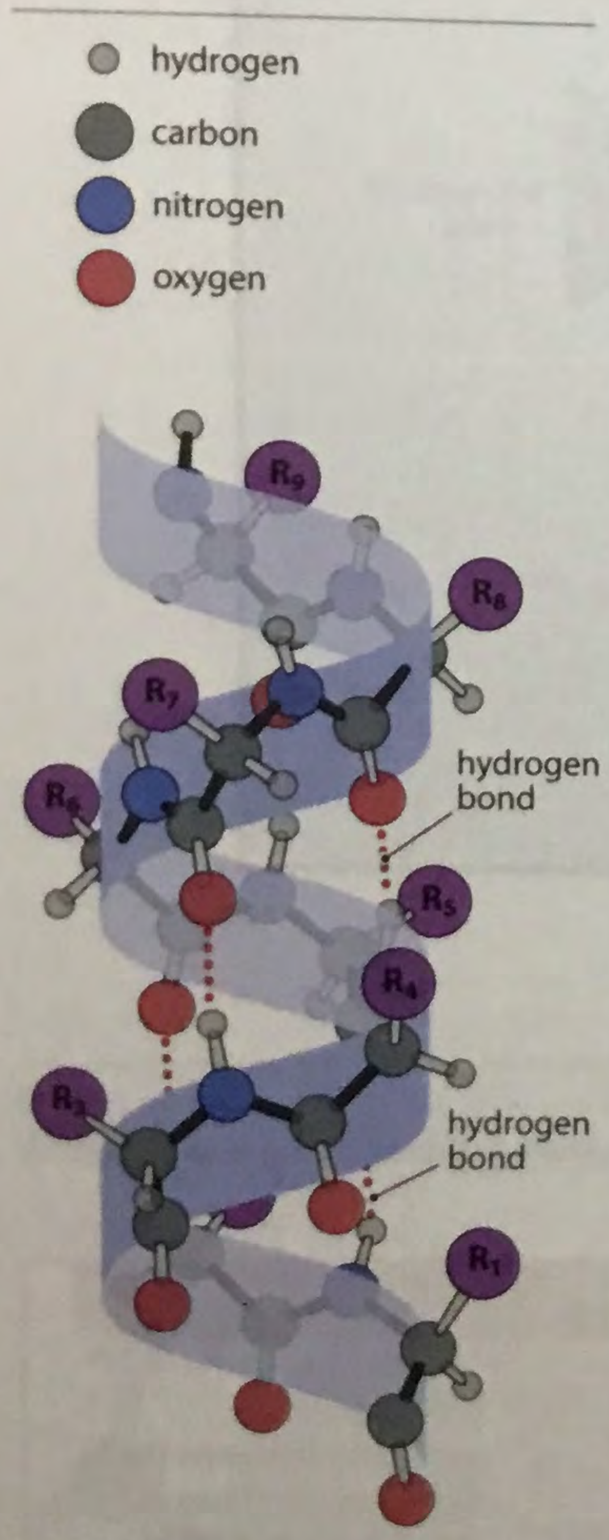


Figure 3.16 The α-helix.
The polypeptide chain is shown in outline. Hydrogen bonds occur between the CO of one peptide group and the NH of the peptide group four positions along the polypeptide.

Other amino acids are 'helix breakers' and either prevent a helix from forming or limit the length of one that does form. Proline is the main example of a helix breaker, the structure of its unusual side-chain (see *Fig. 3.3*) not allowing rotation about the N–C_α bond. The *phi* angle next to a proline is therefore invariant and cannot adopt the degree of rotation needed to form the α-helix. Often a proline is found at one or other end of an α-helix, marking the point where helix formation should terminate.

The β-sheet is another common secondary structure

The β-sheet was also predicted by Pauling and his colleagues following model-building experiments. As with an α-helix, a β-sheet is held together by hydrogen bonds between the CO and NH parts of different peptide groups. The distinction is that in a β-sheet the peptide groups that participate in hydrogen bonding are not close to one another in the polypeptide. In fact, their distance apart is immaterial. What is

Box 3.5 What is the difference between a left-handed and a right-handed helix?

The easiest way to answer this question is to imagine that the helix is a spiral staircase and you are climbing up that staircase. If the helix is right-handed then you will hold the outside rail with your right hand. If it is a left-handed helix then the outer rail will be adjacent to your left hand. The famous spiral staircase at the Loretto Chapel in Santa Fe, New Mexico, which is claimed to have been constructed miraculously by St Joseph, is a left-handed helix. The equally famous spiral staircase at the Vatican Museums, designed by Giuseppe Momo in 1932, is right-handed.

important is that a series of hydrogen bonds forms between two parts of a polypeptide so that those segments are held together side by side (*Fig. 3.17*). Addition of more segments results in a sheet-like structure that can comprise 10 or more strands, each containing up to 15 amino acids. Within a **parallel β-sheet**, all the strands run in the same direction (N→C or C→N), whereas in the **antiparallel** version, adjacent strands run in opposite directions. A mixture of the two is also possible in a single sheet. The sheet itself might exhibit some degree of curvature in the form of a right-handed twist.

Stable hydrogen bonds will not form between adjacent strands if the polypeptides are fully extended, where the two rotation angles *psi* and *phi* are both 180°. Instead, there has to be some rotation of the bonds around the α-carbons so that *psi* is about 113° and *phi* about –119° in a parallel sheet, or 135° and –139° in an antiparallel one. These rotations give the polypeptide a zigzag shape and a sheet a pleated appearance (see *Fig. 3.17*). The side-chains point outwards at right angles to the plane of the sheet.

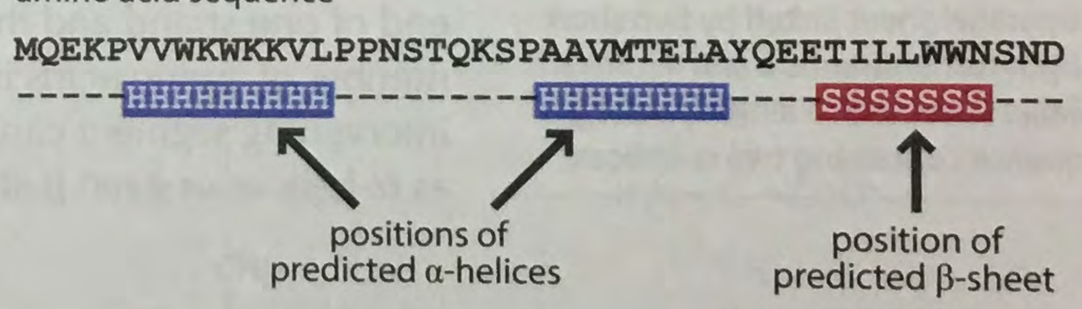
There is little interaction between the side-chains of different amino acids in individual or separate strands which means that, unlike with an α-helix, there are few rules regarding which amino acids can or cannot participate in a β-sheet. Proline is

RESEARCH HIGHLIGHT

Box 3.6 Predicting the secondary structure of a polypeptide from its amino acid sequence

It is easier to work out the amino acid sequence of a protein than its three-dimensional structure, especially as an amino acid sequence can be deduced from the DNA sequence of a gene, using the rules of the genetic code. DNA sequencing is relatively easy, as we will discover in *Section 19.2*, whereas the methods needed to determine the three-dimensional structure of a protein, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy (*Section 18.1.3*), are more difficult and time-consuming. This means that there are a substantial number of proteins whose amino acid sequences are known but their three-dimensional structures are uncharacterized. In the cell the amino acid sequence specifies the three-dimensional structure of the protein. So is there any way in which we can predict that three-dimensional structure simply by examining the amino acid sequence?

Biochemists have attempted to devise rules for predicting protein structure since the 1960s. The early methods concentrated on trying to deduce the positions of α-helices in a polypeptide chain, making use of theoretical information about which amino acids should be helix formers and which helix breakers, along with actual knowledge of the frequencies at which different amino acids were present in helices in the small number of proteins whose structures were actually known at that time. In this way it proved possible to identify the positions of α-helices with 60–70% accuracy. A similar approach allowed β-sheets to be identified with slightly less confidence.



These methods only enable the secondary structure of a protein to be deduced with any degree of certainty. Predicting the way in which the polypeptide, containing its α-helices and β-sheets, folds up into its three-dimensional tertiary structure proved much more difficult. Gradually, the number of proteins whose structures were known increased to the stage where comparisons between the actual structures of related proteins with similar sequences could be made. Then it became possible to devise computer programs that would compare a new amino acid sequence with all the known protein structures, identify entire proteins or parts of proteins with similar sequences, and then use the structures of those proteins to predict the structure taken up by the new amino acid sequence. Even today this method is still not entirely accurate, but it provides a rapid way of identifying the important structural features of a protein before the results of a full X-ray crystallography or NMR analysis are obtained.

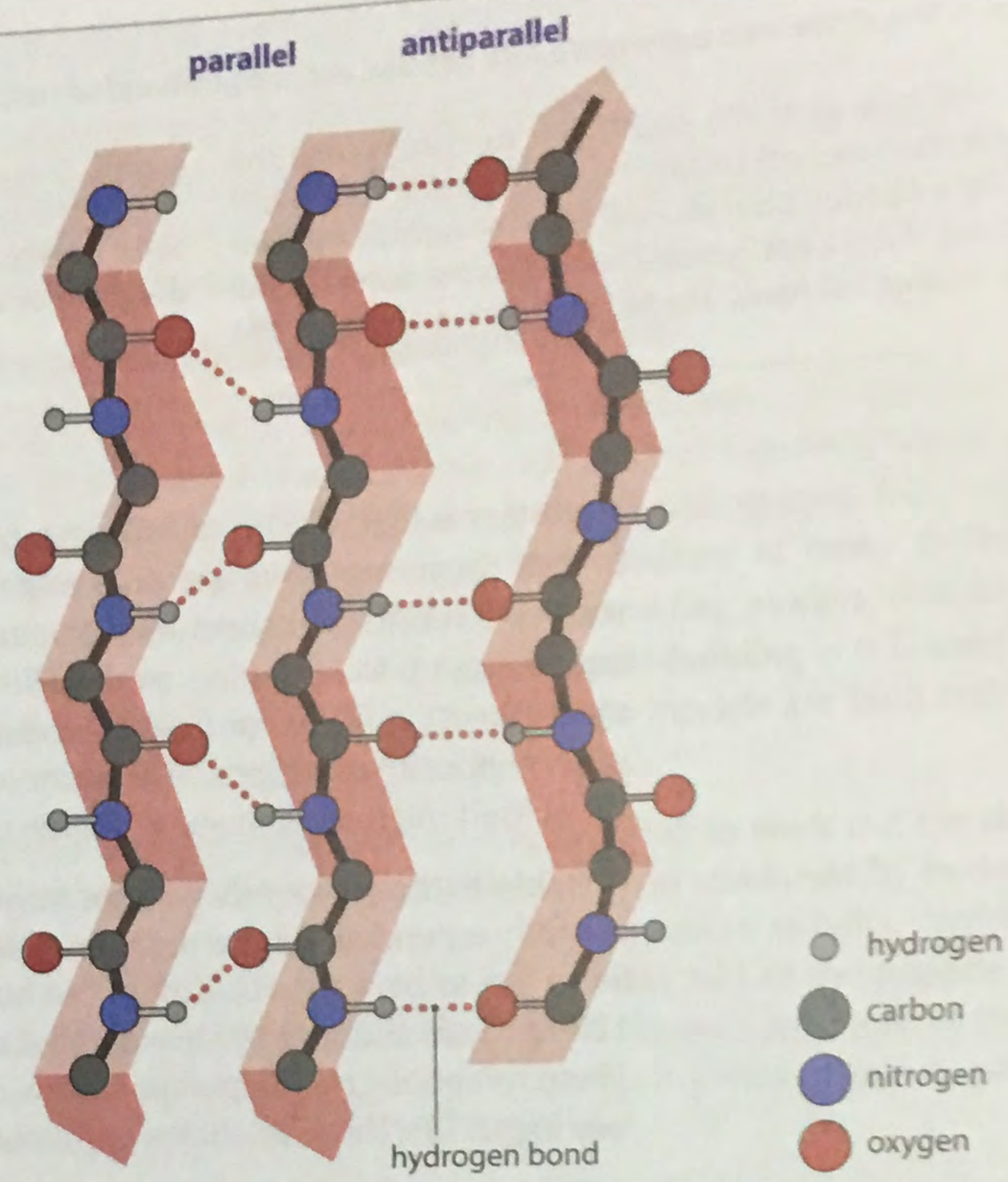


Figure 3.17 The β-sheet.
The polypeptide chains are shown in outline with the R groups omitted. The right-hand and middle strands form an antiparallel β-sheet, in which the polypeptides run in opposite directions. The middle and left-hand strands form a parallel sheet. Note the pleated appearance of the sheets.

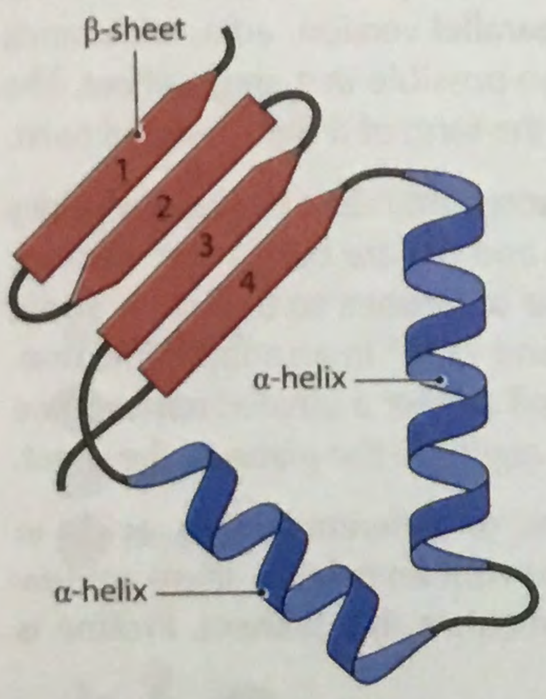


Figure 3.18 A typical combination of β-sheet and α-helices.
In this example there is a four-strand β-sheet. Strands 1–3 form an antiparallel sheet linked by two short hairpin turns. Strands 3 and 4 form a parallel sheet with a lengthy linking sequence containing two α-helices.

once again unfavored and if present is likely to be restricted to one of the edge strands, and amino acids with larger side-chains tend to be located towards the middle of the sheet. There are also very few rules regarding the number of amino acids between the end of one strand and the start of the next. The minimum is usually four because this number of amino acids is needed to execute a hairpin turn of the polypeptide. But the intervening segment can be much longer and can contain other structural motifs such as α-helices or even β-strands participating in a second, separate sheet (*Fig. 3.18*).

3.3 Fibrous and globular proteins

Proteins can broadly be divided into two types, **fibrous** and **globular**. Globular proteins are often soluble and perform a multitude of different functions in living cells. When we study their structure in the next section we will see that they comprise α-helices and β-sheets folded into complex three-dimensional **tertiary structures**. Fibrous proteins are insoluble and usually have more specialized structural roles. These proteins are not folded into tertiary structures. Instead the secondary structure is their highest level of organization.

3.3.1 Fibrous proteins: keratin, collagen and silk

Keratin, collagen and silk are three examples of fibrous proteins. Keratin is present in the hair, horns, nails and skin of animals. The protein is made up of two polypeptides, each of which is composed almost entirely of a slightly compacted version of the α-helix, with 3.5 rather than 3.6 amino acids per turn. This compaction gives the right-handed α-helix a left-handed **superhelix** conformation (*Fig. 3.19*). The configuration of

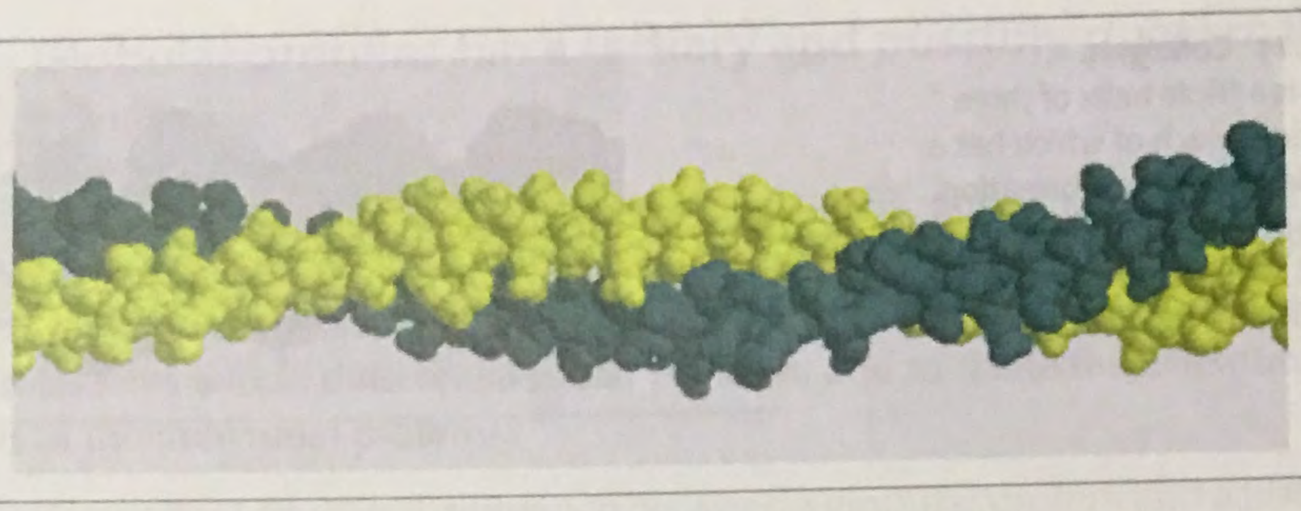
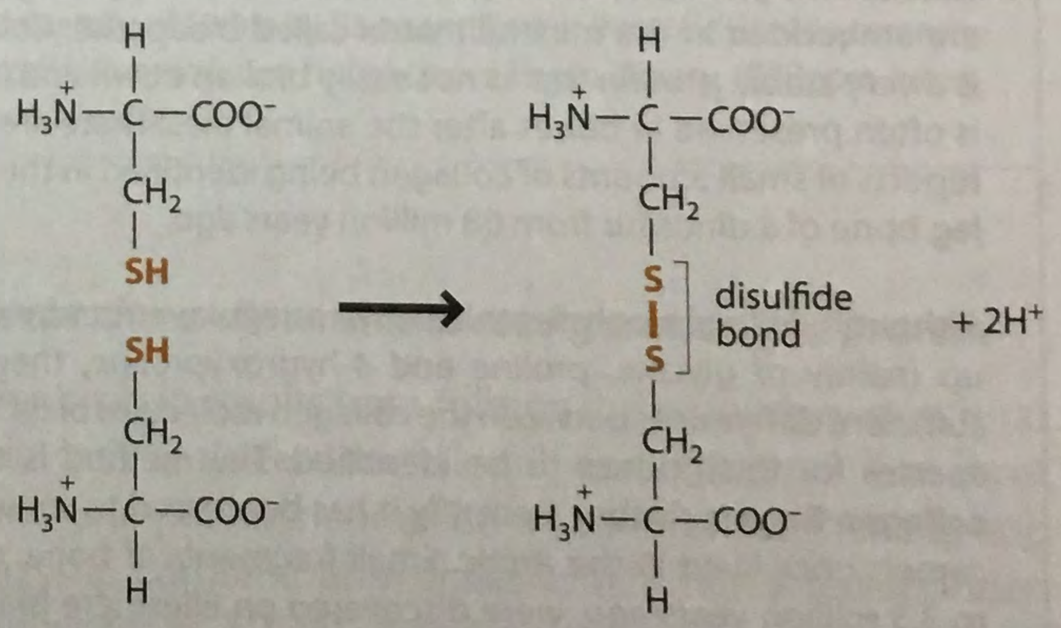


Figure 3.19 Keratin.
The individual polypeptides adopt compacted α-helix structures giving the strand a superhelix conformation. This diagram shows two polypeptides coiled around one another.
Image from *Essential Biochemistry* by Pratt *et al.* with permission from John Wiley and Sons, Inc.

the superhelix is such that two keratin polypeptides can coil around one another, held together by weak bonds called **van der Waals forces**, and possibly by **disulfide bonds**, which are covalent bonds that form between cysteine residues occupying adjacent positions in the two polypeptides (*Fig. 3.20*). The resulting structure, technically called a **coiled coil**, can form fibrils with other coiled coils, which associate further to form microfilaments with high tensile strength, meaning that they are difficult to break by pulling at the ends. The helical conformation of the keratin polypeptide is therefore directly responsible for the physical properties of hair and other structures in which the protein is found.

Collagen also has a helical structure but one that presents new features that we have not encountered so far. A collagen polypeptide has a relatively simple primary structure made up of many repeats of the sequence glycine–X–Y, where X is frequently proline and Y the modified version of proline called 4-hydroxyproline (see *Fig. 3.11*). The repeat is therefore synthesized as glycine–proline–proline, with the second proline in the series converted to 4-hydroxyproline after the polypeptide has been made. The high proline content confers a left-handed helical structure on the polypeptide, with 3.3 amino acids per turn, partly because there cannot be any rotation about the N–C_α bond of a proline, and partly because the proline side-chains repel one another and try to be as far apart as possible. Three of these helical polypeptides then coil around



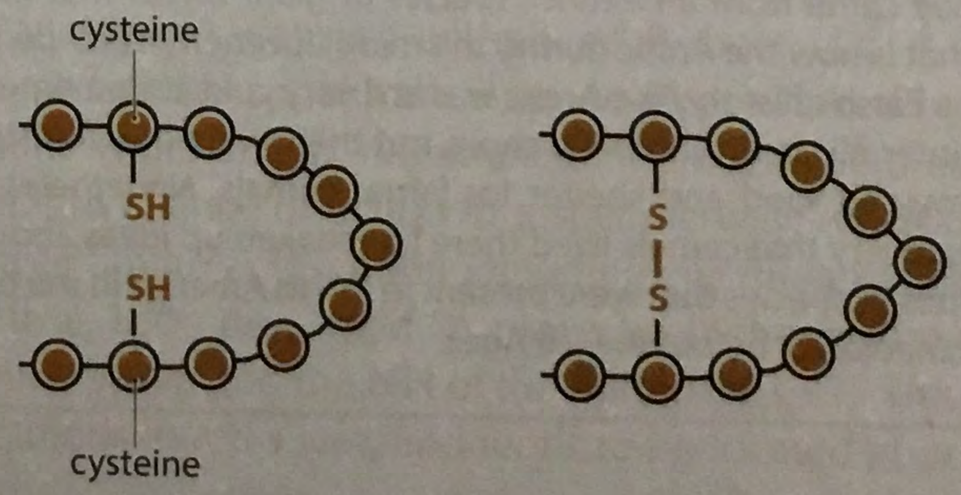


Figure 3.20 Disulfide bonds.
The upper drawing shows the chemical structure of a disulfide bond. Below is the effect that formation of a disulfide bond can have on the structure of a polypeptide.

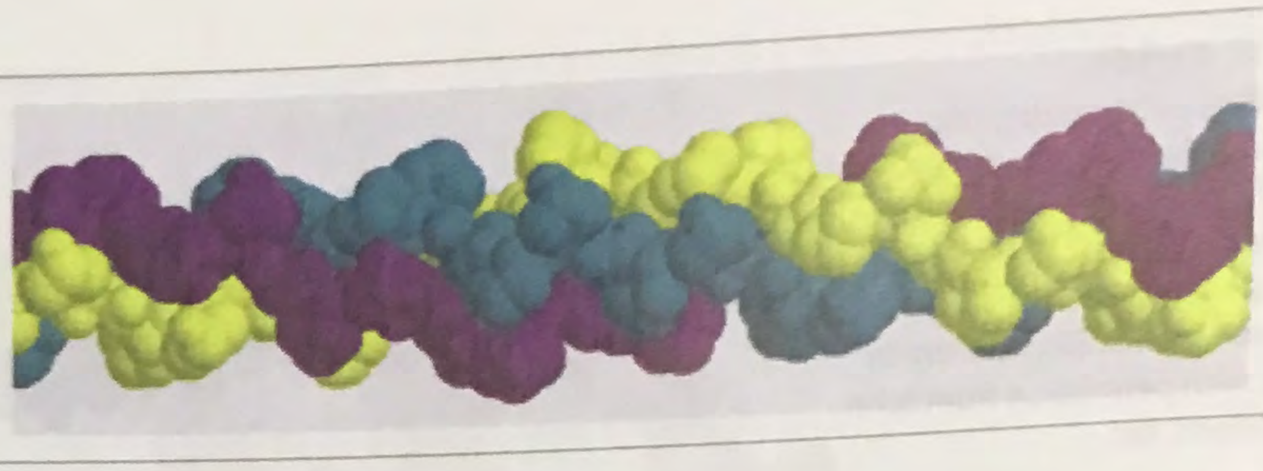


Figure 3.21 Collagen. Collagen is a triple helix of three polypeptides, each of which has a left-handed helical conformation. Image from *Essential Biochemistry* by Pratt *et al.* with permission from John Wiley and Sons, Inc.

one another to make a right-handed **triple helix** (*Fig. 3.21*). The structure requires every third amino acid in each of the polypeptides to be placed close to the central part of the triple helix. This is why every third amino acid is glycine. This amino acid, with its very small side-chain, is the only one that can fit in. The triple helix is held together by hydrogen bonding between the NH of a glycine in one polypeptide and the CO of a peptide group in one of the other two polypeptides. As in keratin, groups of collagen triple helices come together to form fibrils that provide collagen with the strength it needs to play its structural role in connective tissues including bones and tendons.

Silk is rather different. This fiber is produced by various insects and exploited by humans to make fine fabrics. The fibrous component of silk is the protein called fibroin, but this does not have a helical structure. Instead, each fibroin polypeptide has a high glycine and alanine content, which enables it to form extensive β-sheets, which layer on top of one another, with very close packing because of the small size of the glycine and alanine side-chains. The individual β-sheets provide tensile strength, but the layers of sheets are held together less tightly. This means that silk fibers are both strong and flexible.

Box 3.7 Using collagen structure to identify extinct animals RESEARCH HIGHLIGHT

Collagen is one of the most important proteins in vertebrates, being present in bones, tendons and other structural tissues. In bones, collagen fibrils make up about 20% of the dry weight and are embedded in the mineral matrix called bioapatite. Collagen is a very stable protein that is not easily broken down and hence is often preserved in bones after the animal dies. There are even reports of small amounts of collagen being identified in the fossil leg bone of a dinosaur from 68 million years ago.

Although collagen polypeptides have a regular structure made up mainly of glycine, proline and 4-hydroxyproline, there are sufficient differences between the collagen molecules of different species for fossil bones to be identified. The method is called **collagen fingerprinting**. Recently it has been used to show that camels once lived in the Arctic. Small fragments of bone, dated to 3.5 million years ago, were discovered on Ellesmere Island in the Canadian High Arctic. Collagen fingerprinting showed that they came from an extinct species of giant camel that lived in what is now the Arctic during the mid-Pliocene, a warm period in the Earth's history. The Arctic was still very cold at that time, with winter blizzards and deep snow, but there were also forests that provided food and shelter for large animals. Nonetheless, the discovery that camels lived there has shaken up ideas about the types of species that were present in North America in the period immediately before the Ice Ages.

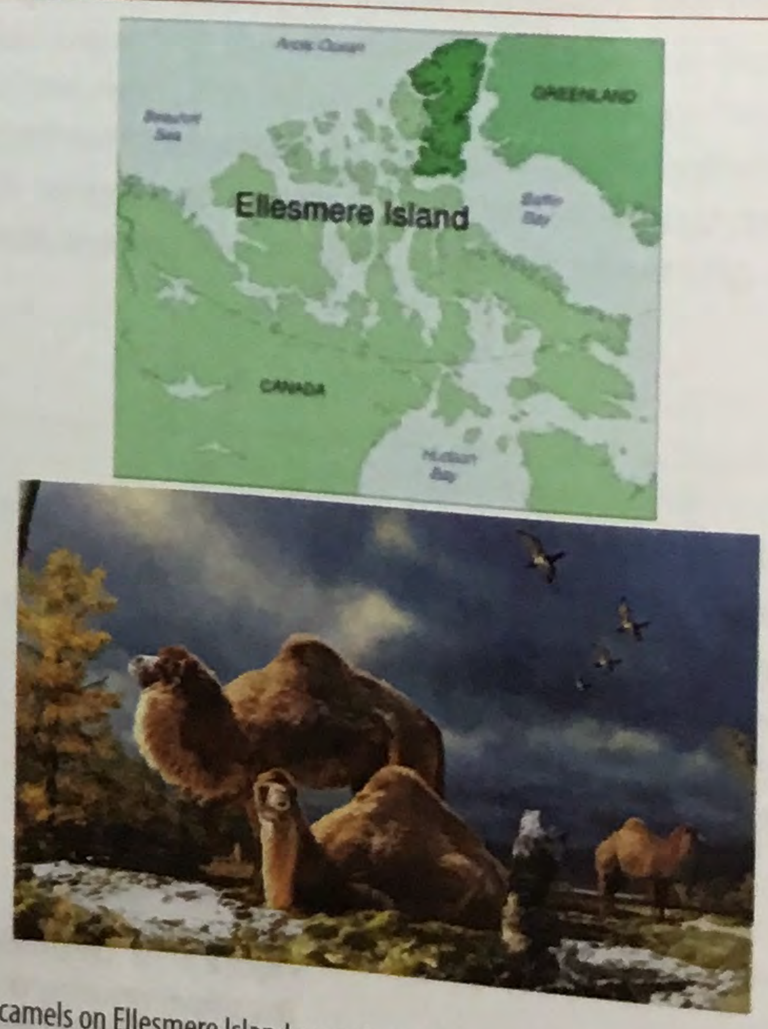


Image of camels on Ellesmere Island reproduced with permission from the artist Julius Csotonyi.

3.3.2 Globular proteins have tertiary and possibly quaternary structures

Globular proteins have spherical rather than elongated fibrous structures, and most are soluble in water. They have diverse biochemical roles, and equally diverse structures. In fact a major goal of biochemistry over the last 20 years has been to identify common structural features within different globular proteins, and to relate those features to the functions of the individual proteins.

The most important structural difference between a globular protein and a fibrous one is that the former displays at least one, and possibly two, higher levels of organization. These are called the **tertiary** and **quaternary structures**, and understanding them is the key to understanding the importance of globular proteins in biochemistry.

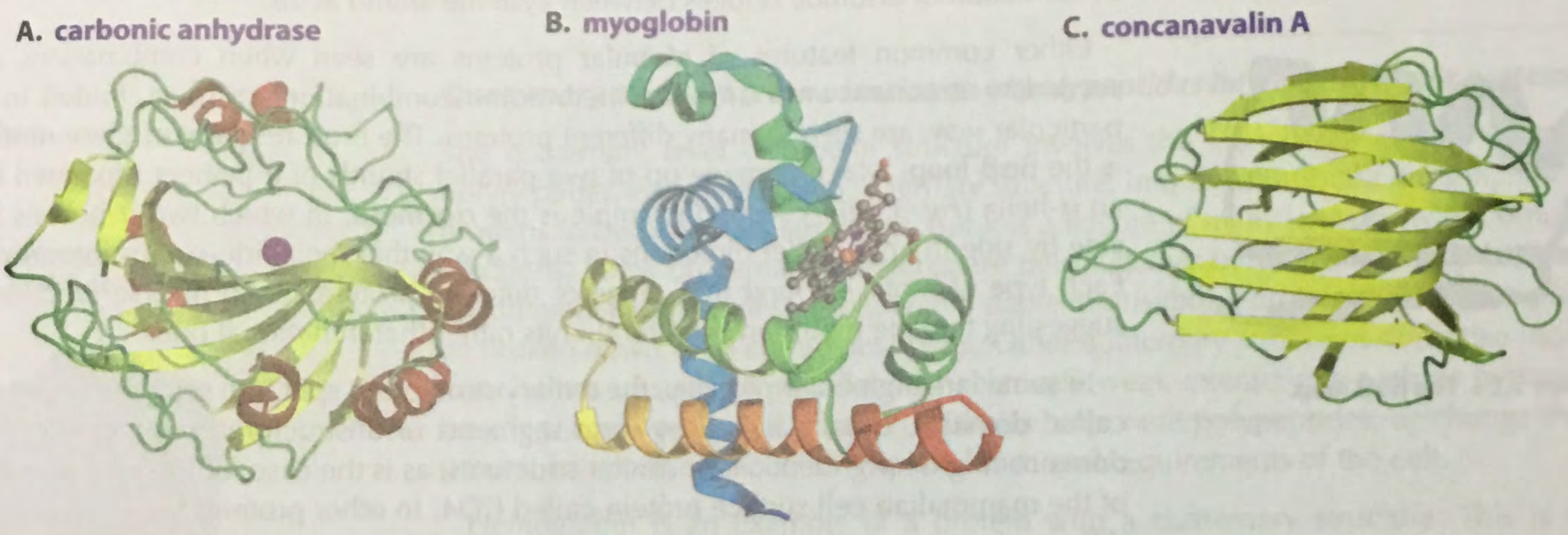
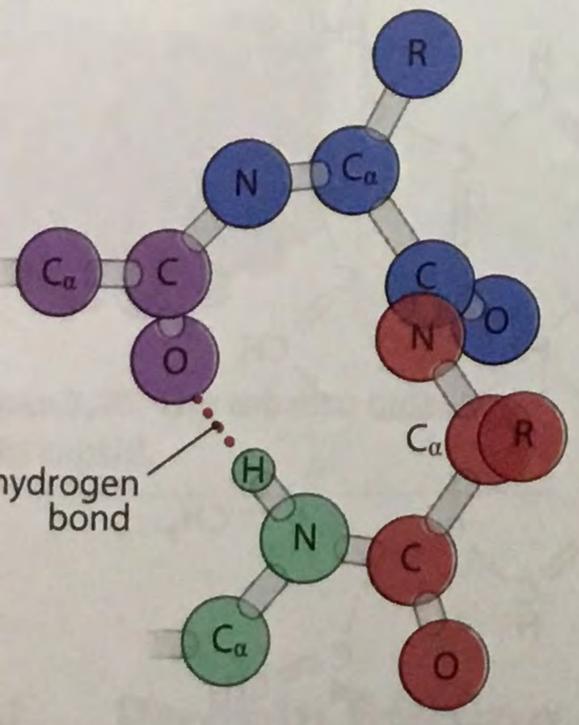


Figure 3.22 Three globular proteins.
(A) Carbonic anhydrase. This protein comprises a ten-stranded β-sheet (yellow) surrounded by five α-helices (pink and red). It also contains a zinc atom, shown in blue. (B) Myoglobin. The secondary structures are all α-helices. It also contains a heme molecule. (C) Concanavalin A. The secondary structure is entirely β-sheet.
(A) Reproduced with permission from University of Maine by Raymond Fort Jr (http://chemistry.umeche.maine.edu/CHY431.html). (B) Reproduced with permission from Science Photo Library. (C) Reproduced from Wikipedia under a Creative Commons license.



ure 3.23 A β-turn.
e four amino acids involved in the n are shown in different colors. For ity, the hydrogen atoms are left , except for the one that ticipates in the hydrogen bond.

The tertiary level of structure is the three-dimensional configuration of a protein

The tertiary structure of a globular protein results from folding the secondary structural components of the polypeptide into a three-dimensional configuration. For most proteins, the secondary structural components comprise a mixture of α-helices and β-sheets. An example is the enzyme carbonic anhydrase, which has a ten-stranded β-sheet surrounded by five α-helices (*Fig. 3.22A*). A few proteins have more uniform tertiary structures. Myoglobin, for example, is made up of eight α-helices and no β-sheet (*Fig. 3.22B*), and concanavalin A consists entirely of β-sheet (*Fig. 3.22C*). Whatever the combination, the secondary structural components are linked by less organized segments of polypeptide which might, nonetheless, include structures that cause the path of the polypeptide to change direction in a specific way. An example is the β-turn, which comprises four amino acids, often including glycine and proline. These four amino acids execute a 180° turn, held in place by hydrogen bonding between the CO of the first peptide group and the NH of the third (*Fig. 3.23*). This type of turn often connects pairs of strands in a β-sheet, and tends to be located at or near the surface of a globular protein.

Although the structures of globular proteins are very diverse, certain common features can be identified. The most consistent of these features is the distribution of the 'water-fearing' and 'water-loving' parts of the polypeptide chain. In most globular proteins, all of the nonpolar amino acid side-chains are located inside the structure. This is exactly what we expect because these side-chains are hydrophobic and so will tend to become buried within the protein when it folds into its tertiary structure. Similarly, the polar and charged parts of the polypeptide will usually be on the surface so that they can make contact with water molecules, presuming the protein is in an aqueous environment such as the inside of a cell. The polar parts of a polypeptide include not just amino acid side-chains, but also the CO and NH groups of the peptide linkages, unless these have formed hydrogen bonds, as in an α-helix or β-sheet. Peptide groups that are not participating in hydrogen bonding will therefore tend to be on the surface of the protein. These various forces dictate the way in which the polypeptide folds into its tertiary structure. Once folded, the structure will be stabilized by various interactions, such as van der Waals forces, and possibly by formation of disulfide bridges between cysteine amino acids.

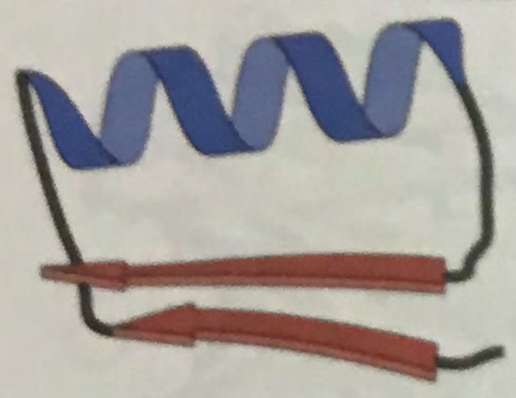


Figure 3.24 The βαβ loop.

Other common features of globular proteins are seen when combinations of secondary structural units are examined. Some combinations of units, folded in a particular way, are seen in many different proteins. The most frequent of these **motifs** is the **βαβ loop**, which is made up of two parallel strands of a β-sheet separated by an α-helix (*Fig. 3.24*). A second example is the **αα motif**, in which two α-helices lie side by side in antiparallel directions in such a way that their side-chains intermesh. Each type of motif is found in a range of different proteins, with diverse functions, suggesting that the motifs are structural units rather than functional ones.

In some larger globular proteins, the tertiary structure is split into separate segments called **domains**, usually linked by short segments of unstructured polypeptide. The domains might have identical or similar structures, as is the case for the four domains of the mammalian cell surface protein called CD4. In other proteins the domains are different in structure, each possibly contributing a different part of the overall function of the protein.

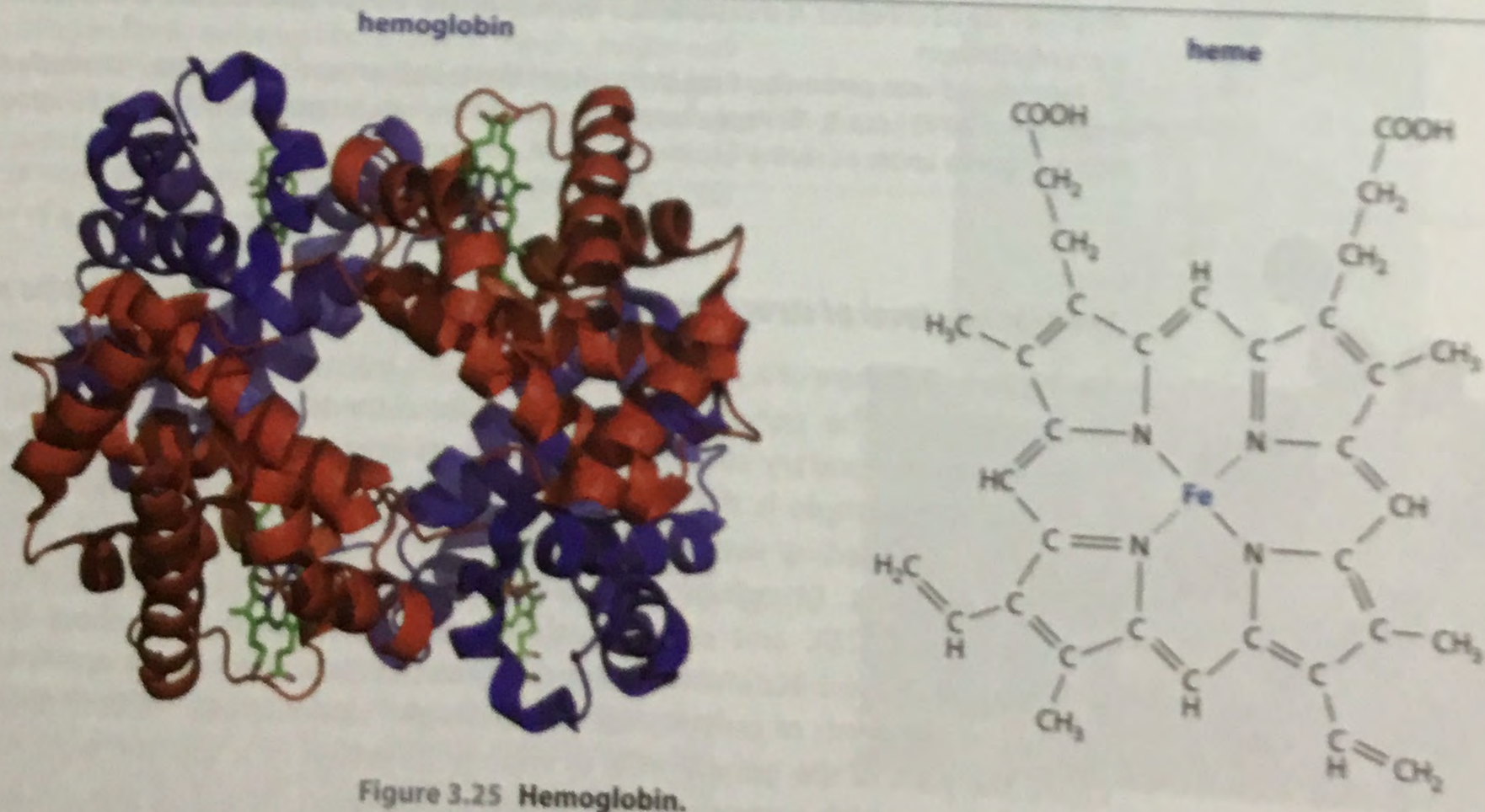


Figure 3.25 Hemoglobin.
This protein is a tetramer of two identical α subunits and two identical β subunits. The heme groups are shown in green in the protein structure. Heme is an organic compound containing an iron atom, which reversibly binds oxygen, enabling hemoglobin in red blood cells to carry oxygen from the lungs to other parts of the body.
(A) Haemoglobin image by Zephyris reproduced from Wikipedia under a CC BY-SA license.

Box 3.8 An example of a protein with a mixture of domains

Human tissue plasminogen activator (TPA), which is involved in blood clotting, is a good example of a multidomain protein. TPA has five domains:

- Two identical 'kringle' structures, which enable TPA to bind to other proteins, and also to lipids that act as mediators in the blood clotting process. Each kringle structure is a large loop stabilized by three disulfide bonds.
- A 'finger' module, which is a small β-sheet structure that binds fibrin, a fibrous protein found in blood clots.
- A growth factor module, made up of three loops held in place by two disulfide bonds. This module enables TPA to stimulate cell proliferation as part of the wound healing response.
- A large protease domain, comprising a β-sheet and α-helix. The function of the protease domain is to convert an inactive protein called plasminogen into its active form, called plasmin. The protease does this by cutting a single peptide bond within the plasminogen polypeptide. Plasmin breaks down unused fibrin, ensuring that the clot does not spread into the bloodstream.

All of these domains are also found, with very similar structures, in other proteins. Kringle and finger domains are common in proteins involved in blood clotting, and growth factor domains are found in several proteins that stimulate cell growth. One of these, epidermal growth factor, is made up simply of a single growth factor domain.

Quaternary structure is the association of polypeptides into multisubunit proteins

The quaternary level of protein structure involves the association of two or more polypeptides, each folded into its tertiary structure, into a multisubunit protein. Not all proteins form quaternary structures, but it is a feature of many proteins with complex functions. Some quaternary structures are held together by disulfide bonds between the different polypeptides, resulting in a stable multisubunit protein that cannot easily be broken down to its component parts. Other quaternary structures comprise looser associations of subunits stabilized by relatively weak interactions such as hydrogen bonding. These proteins can revert to their component polypeptides, or change their subunit composition, according to the functional requirements of the cell.

Hemoglobin is an example of a protein with a quaternary structure. This is the protein in vertebrate red blood cells that carries oxygen from the lungs to other tissues in the body. It is a tetramer of four polypeptides, comprising two identical α subunits and two identical β subunits (*Fig. 3.25*). The polypeptides are called globins so the subunits are α-globins and β-globins. Each globin has an attached heme group, a non-protein compound that binds oxygen. The quaternary structure is stabilized by hydrogen and electrostatic bonds between the globin subunits.

Large quaternary structures are formed by the proteins that make up the coats, or capsids, of viruses. The capsid of tobacco mosaic virus (TMV), for example, is made up of 2130 identical subunits. Each subunit is a small globular protein, comprising 158 amino acids folded into a tertiary structure that includes four α-helices. The subunits are arranged into a tightly packed helical structure with 16.3 subunits per turn, which encloses the RNA genome of the virus. The TMV capsid is, in effect, a single multisubunit quaternary protein (*Fig. 3.26*). TMV is an example of a filamentous virus but the same principle applies to the capsids of icosahedral viruses. Human poliovirus has an icosahedral capsid with 20 faces. Each face is made up of 12 polypeptide subunits, three each of VP1, VP2, VP3 and VP4. The capsid as a whole therefore has 240 subunits, 60 of each of the four VP subunits.

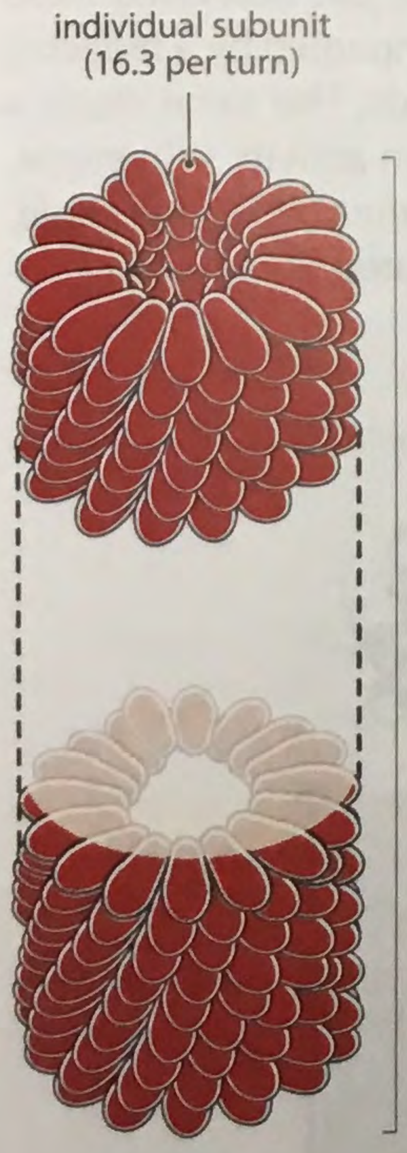


Figure 3.26 The tobacco mosaic virus capsid.

3.4 Protein folding

A fundamental notion regarding globular proteins is that the secondary and tertiary structures are specified by the amino acid sequence of the polypeptide. In other words, a particular amino acid sequence will fold into just one, and no other, tertiary structure. This was first demonstrated by experiments carried out in the 1950s and has

led to detailed models for the folding process and for the role in the cell of **molecular chaperones**, proteins that help other proteins to fold.

3.4.1 Small proteins fold spontaneously into their correct tertiary structures

The notion that the amino acid sequence contains all the information needed to fold the polypeptide into its correct tertiary structure derives from experiments carried out by Christian Anfinsen in the 1950s. He worked with ribonuclease, a small protein of 124 amino acids, whose tertiary structure is a mixture of α-helices and β-sheet and includes four disulfide bonds between cysteine amino acids in different parts of the polypeptide. Anfinsen used ribonuclease that had been purified from cow pancreas and resuspended in an aqueous buffer. The addition of urea, a compound that disrupts hydrogen bonding, resulted in a decrease in the activity of the enzyme, measured by testing its ability to cut up molecules of RNA into their monomeric subunits (*Fig. 3.27*). At the same time, the viscosity of the solution increased, indicating that the protein was being **denatured** by unfolding into an unstructured polypeptide chain.

The urea was then removed from the solution by **dialysis**. The viscosity decreased and the protein gradually regained its ability to cut RNA. The protein therefore refolds spontaneously when the denaturant is removed.

Urea does not disrupt disulfide bonds, so in the experiment just described these bonds remain intact. In a second experiment the urea was accompanied by a reducing agent, β-mercaptoethanol, which does break the disulfide bonds. The same result is obtained when the urea is removed from the solution: the protein activity still returns. This shows that the disulfide bonds are not critical to the protein's ability to refold, they merely stabilize the tertiary structure once it has been adopted.

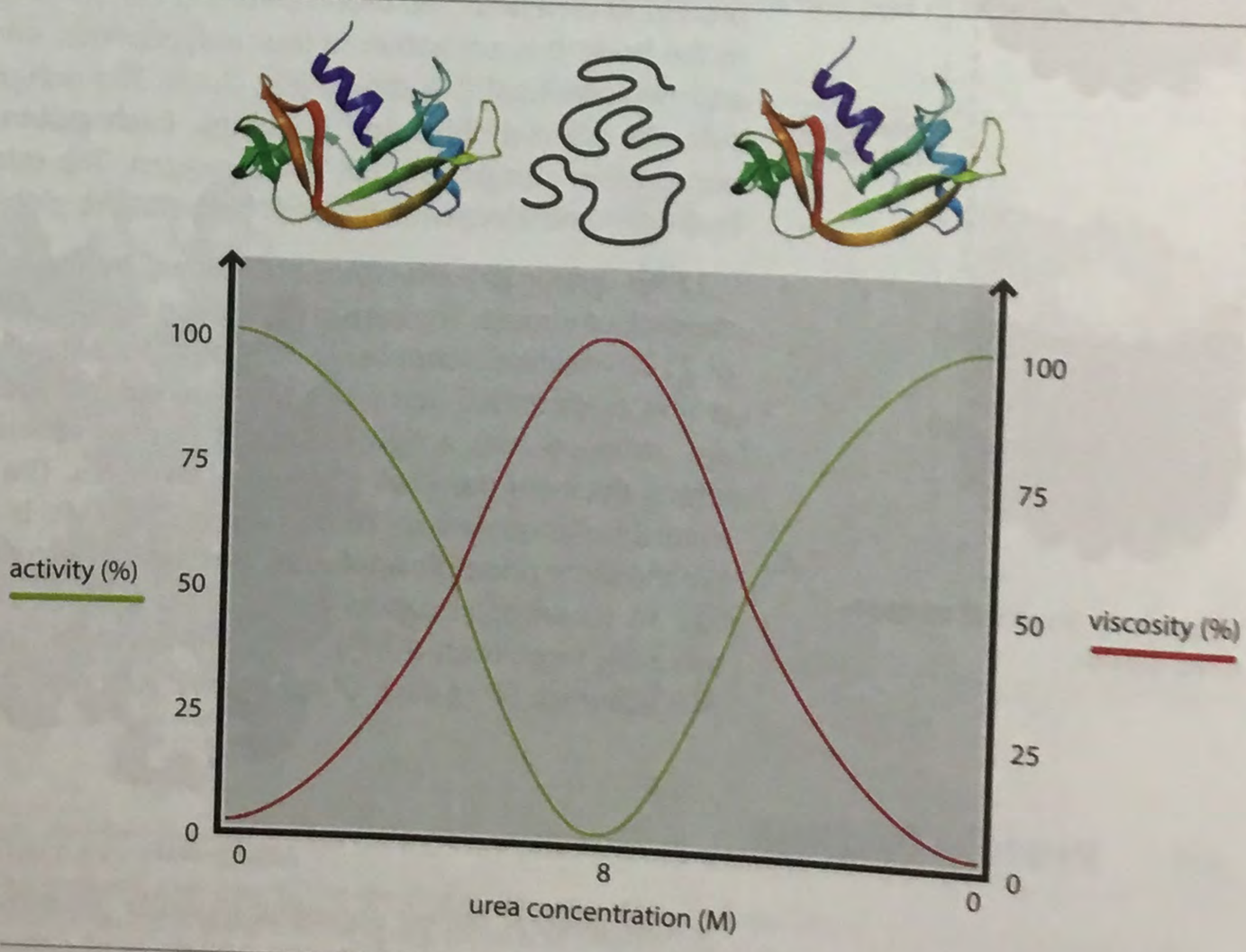


Figure 3.27 Denaturation and spontaneous renaturation of ribonuclease.
The graph shows the changes in ribonuclease activity and solution viscosity that occur as the urea concentration is increased or decreased. As the urea concentration increases to 8 M, the protein becomes denatured by unfolding. Its activity decreases and the viscosity of the solution increases. When the urea is removed by dialysis, this small protein readopts its folded conformation. The activity of the protein increases back to the original level and the viscosity of the solution decreases.
Ribonuclease structure images reproduced from Wikipedia under a CC BY-SA 2.5 license.

3.4.2 Protein folding pathways

Once it had become clear that proteins adopt their tertiary structures spontaneously, biochemists turned their attention to the folding process itself. It was quickly realized that the process cannot be random. It is not possible for a protein simply to explore all the possible conformations that it can take up until it finally hits on the correct one. This was made clear by Cyrus Levinthal in 1969, whose argument is as follows. The tertiary structure of a protein is set by the three-dimensional conformation of the polypeptide. This in turn is set by the *psi* and *phi* values for the bonds on either side of the α-carbons along its polypeptide chain. Remember that it is only by rotation around these bonds that the polypeptide can change direction. Levinthal argued that there would be at least three possible values for each *psi* and *phi* angle, which is almost certainly an under-estimate. This would mean that a polypeptide of 100 amino acids could adopt 3^{198} different conformations – this is about 10^{100}. Even if the protein could explore 10^{13} conformations per second (likely to be an *over*-estimate) this would mean that it would take about 10^{87} seconds for all conformations to be checked. This is a huge amount of time, longer even than the age of the universe (in fact, much longer). So proteins cannot find their correct tertiary structures just by a random search. This problem has been called **Levinthal's paradox**.

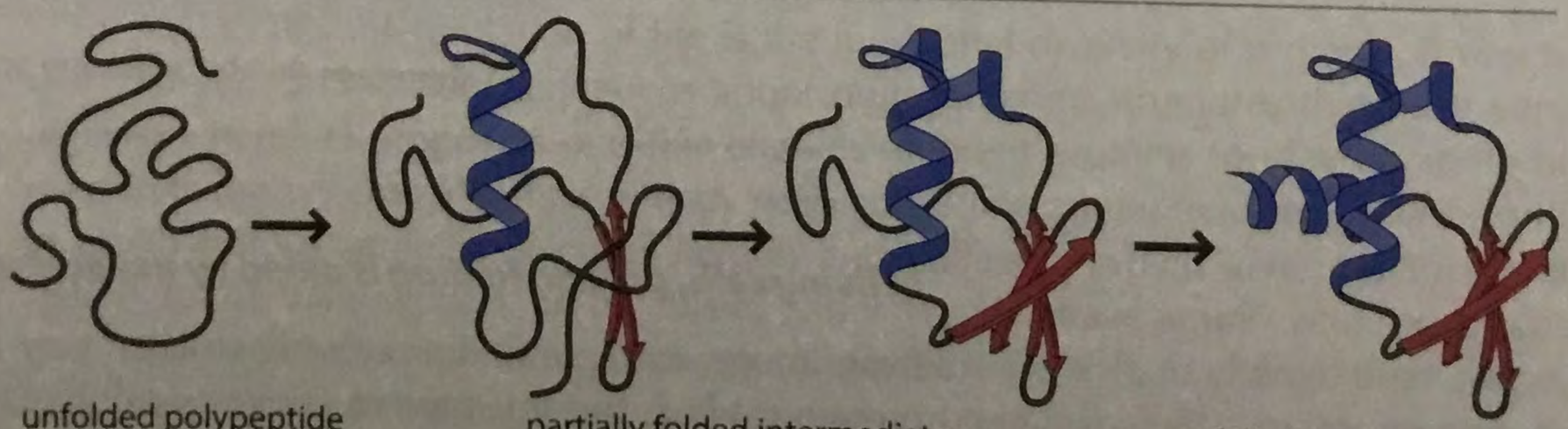
The folding process must be ordered in some way

Levinthal's paradox shows that the folding process must be ordered in some way. This had led biochemists to conclude that there is a **folding pathway** for each protein, each step involving just a small part of the polypeptide (*Fig. 3.28*). In this way the protein could find its way to its correct structure without having to test every possible conformation. These considerations, combined with experimental studies of protein folding, have led to the **molten globule** model. In this model, the initial step in folding is the rapid collapse of the polypeptide into a compact structure, with slightly larger dimensions than the final protein, driven by the desire of the hydrophobic amino acid side-chains to avoid water. Collapse into this molten globule might automatically fold some of the polypeptide into its α-helices and β-sheets. Because the globule is 'molten' it can change conformation rapidly, identifying additional folds so that the correct tertiary structure gradually emerges. For larger proteins, this step might involve construction of correctly folded subdomains which are then brought together to make the final tertiary structure. The whole process can take just a few seconds.

More sophisticated iterations of the molten globule and other models for protein folding imagine a **folding funnel** that the protein passes through, gradually taking up less random conformations until it reaches its final structure (*Fig. 3.29*). As the protein adopts an increasingly folded state, the funnel narrows because there are fewer options for the next steps towards the final structure. There are also side funnels into which the protein can be diverted, leading to an incorrect structure. If an incorrect structure is sufficiently unstable then partial or complete unfolding may occur, allowing the protein to return to the main funnel and pursue a productive route towards its correct conformation.

ermodynamic terms, a
ease in randomness is
mpanied by a reduction in
energy (see *Section 7.2.1*).

.28 **A protein folding**
.



RESEARCH HIGHLIGHT

Box 3.9 Studying protein folding

How do biochemists study the way in which individual proteins fold? Anfinsen's experiments were revolutionary in their day, but that was 60 years ago and all he was able to do was measure the viscosity of his ribonuclease solutions, and the activity of the enzyme, to follow the unfolding and folding of the protein. He could not deduce any specific information about the folding pathway itself.

Biochemists today use three main approaches to study protein folding.

- For some proteins, it is possible to halt the folding pathway at certain points, and then use NMR to study the structure of the intermediate form directly. This provides very specific information about a folding pathway, but so far has only been used with a few proteins, and it may not be generally applicable because halting the pathway is not always possible.
- The degree of folding can be followed in real time by methods such as **circular dichroism**. In principle, this is the same as the approach used by Anfinsen, but with modern variations that make it much more informative. Circular dichroism measures the absorption of polarized light by a protein. Secondary structures such as α-helices and β-sheets absorb polarized light, so circular dichroism measures the rate at which these structures are formed. As with Anfinsen's experiments, this type of research is usually performed with proteins that have been denatured and are gradually reforming their folded structures. However, modern methods enable the renaturation process to be controlled much more carefully, so all the proteins in the solution begin to fold at exactly the same time. This means that the process is synchronized and hence much easier to study. It is even possible to study the folding of a single molecule that is initially held in a linear conformation with an **optical tweezer**, a laser device that can be used to manipulate individual molecules.
- The third approach is to change the amino acid sequence of the protein and see what effect this has on the folding pathway. The change in sequence is usually brought about by introducing a **mutation** into the gene coding for the protein (see *Section 19.1.2*). In this way the stage in the folding pathway at which a particular part of a protein adopts its structure can be identified. For example, imagine that we wish to test if a particular α-helix forms early in a folding pathway. To do this, we would change an amino acid predicted to be crucial for nucleating that helix into one that will prevent the helix from forming. If our hypothesis is correct, and the α-helix is indeed important in the early part of the folding pathway, then the altered protein should be unable to fold beyond that stage.

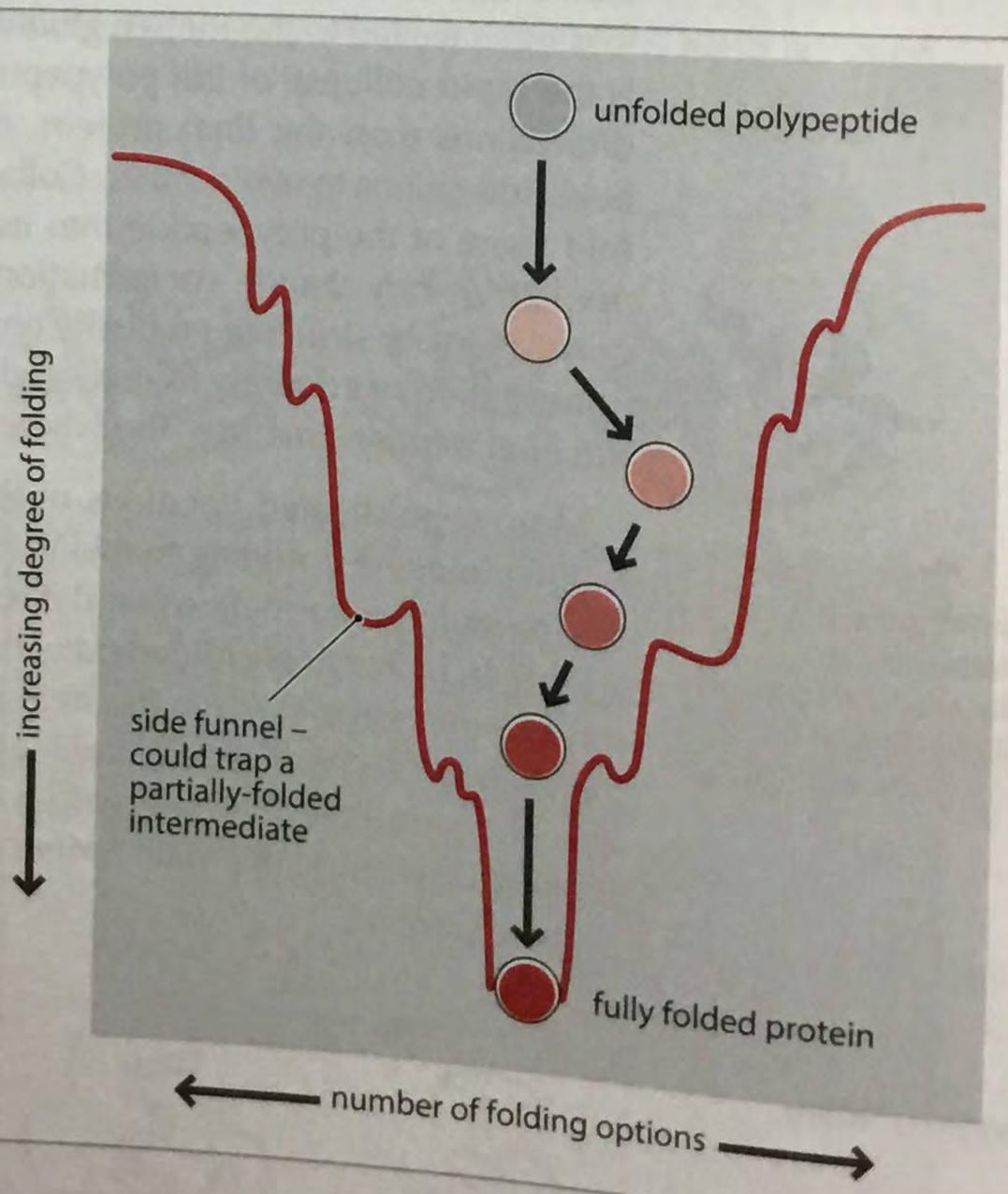


Figure 3.29 A folding funnel. The top of the funnel is wide because the unfolded polypeptide can initially adopt any one of many initial intermediate structures. The funnel gradually narrows as the protein becomes more folded and its options for future folding are reduced. Gradually those options decrease and the protein becomes more completely folded and the fully folded protein emerges from the spout at the bottom. Side funnels lead to dead-ends. If the protein enters one of these side funnels then it has to partially unfold in order to return to the main funnel.

In living cells, protein folding is aided by molecular chaperones

Experiments with purified proteins have been very useful in helping us to understand protein folding, but this type of *in vitro* work has two limitations. First, only smaller proteins with less complex structures fold spontaneously in the test tube. Larger

Other types of proteins have quite different functions. Enzymes are proteins whose amino acid sequences enable them to catalyze biochemical reactions, such as those involved in metabolism. Other proteins have transport functions and carry compounds around the body. We have already studied hemoglobin, which carries oxygen from the lungs to other tissues. A second example is serum albumin, which transports fatty acids, which are the building blocks of lipids and are also used as energy sources.

Some proteins help to store molecules for future use by the organism. Examples include ovalbumin, which stores amino acids in egg white, and ferritin, which stores iron in the liver. A large group of proteins have protective functions, such as the immunoglobulins of mammals, which form complexes with foreign proteins and protect the body against infectious agents such as viruses and bacteria.

There are also **regulatory proteins** that control cellular and physiological activities. These include well-known hormones such as insulin, which regulates glucose metabolism in vertebrates, and the two growth hormones somatostatin and somatotropin. Although made inside a cell hormones are secreted so they can travel around the body and convey their regulatory messages to other cells. Other regulatory proteins work entirely within the cells in which they are synthesized, possibly responding to signals from extracellular hormones. Examples are components of the MAP kinase pathway, which regulate activities such as cell division in response to external signals.

All of these diverse functions are specified by the chemical properties of individual proteins, which in turn are specified by their three-dimensional structures and hence by their amino acid sequences. The adoption of those correct three-dimensional structures by protein folding is therefore one of the fundamental cornerstones of biology.

Further reading

Bragulla HH and Homberger DG (2009) Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. *Journal of Anatomy* **214**, 516–59.

Covington AK, Bates RG and Durst RA (1985) Definition of pH scales, standard reference values, measurement of pH and related terminology. *Pure and Applied Chemistry* **57**, 531–42. ***Everything that you will ever need to know about this subject.***

Eisenberg D (2003) The discovery of the α-helix and β-sheet, the principal structural features of proteins. *Proceedings of the National Academy of Sciences USA* **100**, 11207–10.

Jungck JR (1985) Margaret Oakley Dayhoff, "harnessing the computer revolution". *The American Biology Teacher* **47**, 9–10. ***A review of the work of one of the first bioinformaticians.***

Klug A (1999) The tobacco mosaic virus particle: structure and assembly. *Philosophical Transactions of the Royal Society of London*, series B **354**, 531–5.

Mayer MP (2013) Hsp70 chaperone dynamics and molecular mechanism. *Trends in Biochemical Sciences* **38**, 507–14.

Pauling L and Corey RB (1951) The pleated sheet, a new layer configuration of polypeptide chains. *Proceedings of the National Academy of Sciences USA* **37**, 251. ***The first description of the β-sheet.***

Pauling L, Corey RB and Branson HR (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences USA* **37**, 205–11. ***The first description of the α-helix.***

Ramachandran GN, Ramakrishnan C and Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology* **7**, 95–9. ***Describes the psi and phi angles and the Ramachandran plot.***

Römer L and Scheibel T (2008) The elaborate structure of spider silk. *Prion* **2**, 154–61.

Rost B (2001) Protein secondary structure prediction continues to rise. *Journal of Structural Biology* **134**, 2014–18.

Rybczynski N, Gosse JC, Harington R, Wogelius RA, Hidy AJ and Buckley M (2013) Mid-Pliocene warm-period deposits in the High Arctic yield insight into camel evolution. *Nature Communications* **4**, 1550. ***An Arctic camel identified by collagen fingerprinting.***

Shoulders MD and Raines RT (2009) Collagen structure and stability. *Annual Review of Biochemistry* **78**, 929–58.

Yébenes H, Mesa P, Muñoz IG, Montoya G and Valpoesta JM (2011) Chaperonins: two rings for folding. *Trends in Biochemical Sciences* **36**, 424–32.

Self-assessment questions

Multiple choice questions

Only one answer is correct for each question.
Answers can be found on the website:
www.scionpublishing.com/biochemistry

1. Titin, the longest known polypeptide, has how many amino acids?
- **(a)** 1464
- **(b)** 3685
- **(c)** 21 075
- **(d)** 33 445

2. Which amino acid is given the one-letter abbreviation 'A'?
- **(a)** Alanine
- **(b)** Arginine
- **(c)** Asparagine
- **(d)** Aspartic acid

3. Which amino acid has an unusual side-chain that includes the nitrogen of the amino group attached to the α-carbon?
- **(a)** Asparagine
- **(b)** Proline
- **(c)** Tryptophan
- **(d)** Tyrosine

4. The D- and L-forms of an amino acid are examples of what?
- **(a)** Enantiomers
- **(b)** Isomers
- **(c)** Optical isomers
- **(d)** All of the above

5. What is a molecule that has two ionized groups called?
- **(a)** Enantiomer
- **(b)** Hydrophile
- **(c)** Zwitterion
- **(d)** None of the above

6. Which one of the following statements regarding the isoelectric point of an amino acid is **incorrect**?
- **(a)** It is the pH at which an amino acid has no electrical charge
- **(b)** At the isoelectric point both the carboxyl and amino groups are ionized
- **(c)** It is a pH value greater than the pK_a of the amino group
- **(d)** For glycine, the isoelectric point is just below pH 6.0

7. Which two amino acids have positively charged side-chains at pH 7.4?
- **(a)** Arginine and lysine
- **(b)** Aspartic acid and glutamic acid
- **(c)** Cysteine and tyrosine
- **(d)** Histidine and proline

8. What is the name given to the type of chemical bond that forms between the slightly electropositive hydrogen atom in a polar group and an electronegative atom?
- **(a)** Covalent bond
- **(b)** Electrostatic bond
- **(c)** Hydrogen bond
- **(d)** van der Waals bond

9. Which one of the following statements is a feature of hydrophobic amino acids?
- **(a)** They are readily soluble
- **(b)** They are usually found on the surface of a protein
- **(c)** They have nonpolar side-chains
- **(d)** They often form hydrogen bonds with other hydrophobic amino acids

10. Which one of the following compounds is an example of a modified amino acid that is found in collagen?
- **(a)** 4-hydroxyproline
- **(b)** Pyrrolysine
- **(c)** Selenocysteine
- **(d)** Selenoproline

Short answer questions

These do not require additional reading.

1. Draw the general structure of an amino acid and indicate the chemical groups that participate in formation of peptide bonds.
2. Distinguish between the L- and D-forms of an amino acid and explain how the two configurations are identified experimentally.
3. Define the term pK_a and explain why some amino acids have two pK_a values and others have three. How do these pK_a values affect the chemical properties of different amino acids?
4. Describe the differences between covalent, electrostatic and hydrogen bonds.
5. Explain how the Ramachandran plot enables combinations of the *psi* and *phi* bond angles that give rise to different polypeptide configurations to be identified.
6. Distinguish between the structures of the α-helix and β-sheet.
7. Describe the structure of the fibroin protein and explain how this structure enables silk to be both strong and flexible. To what extent is the structure of fibroin typical of that of other fibrous proteins?
8. Using examples, explain what is meant by the terms 'tertiary' and 'quaternary' with regard to the structure of a globular protein.
9. Summarize the molten globule model for protein folding.
10. Describe the roles of Hsp70 proteins and chaperonins in protein folding.

Self-study questions

These questions will require calculation, additional reading and/or internet research.

1. The Henderson–Hasselbalch equation defines the relationship between pH and pK_a as:

$$pH = pK_a + \log \frac{[A^-]}{[HA]}$$

where $[A^-]$ and [HA] are the concentrations of the ionized and non-ionized forms of a chemical group, respectively. Explain how the Henderson–Hasselbalch equation relates to the ionization graph for glycine shown in *Fig. 3.7*.

2. Draw a graph showing the relative amounts of the different ionized versions of arginine at different pH values. The relevant pK_a values are 2.01 for the carboxyl group, 9.04 for the amino group, and 12.48 for the side-chain.
3. Most proteins denature at temperatures above approximately 50°C because of the disruptive effects that heat has on the chemical bonds that stabilize secondary and tertiary structures. However, some bacteria live at high temperatures, for example in hot springs, and their proteins retain their tertiary structures at temperatures up to 90°C. Speculate on the nature of the structural innovations that might enable a protein to withstand such high temperatures.
4. A protein with a molecular mass of 380 kDa is treated with β-mercaptoethanol. The molecular mass is measured again and now found to be 190 kDa. Provide an explanation for these results.
5. Does the existence of molecular chaperones contradict the statement that the amino acid sequence of a polypeptide contains the information needed to fold that polypeptide into its correct tertiary structure?

11. Which one of the following statements regarding a peptide bond is **incorrect**?
- **(a)** A peptide bond is able to rotate
- **(b)** A peptide bond is formed by a condensation reaction
- **(c)** A peptide bond forms between the carboxyl and amino groups of adjacent amino acids
- **(d)** A peptide bond is a single bond but due to resonance has some double bond characteristics

12. Because of steric effects, what proportion of the possible combinations of *psi* and *phi* bond angles never occur?
- **(a)** 7%
- **(b)** 57%
- **(c)** 77%
- **(d)** All combinations of *psi* and *phi* are possible

13. An α-helix is stabilized by what type of interactions?
- **(a)** Covalent bonds between cysteine amino acids
- **(b)** Hydrogen bonds between complementary amino acids
- **(c)** Hydrogen bonds between peptide groups four positions along the polypeptide
- **(d)** Hydrophobic interactions between peptide groups four positions along the polypeptide

14. A β-sheet is stabilized by what type of interactions?
- **(a)** Covalent bonds between proline amino acids which mark the start and end points of the β-sheet
- **(b)** Hydrogen bonds between complementary amino acids
- **(c)** Hydrogen bonds between two parts of a polypeptide so that those segments are held together side by side
- **(d)** Hydrophobic interactions between different parts of the β-sheet

15. A collagen polypeptide forms what type of secondary structure?
- **(a)** α-helix
- **(b)** β-sheet
- **(c)** Double helix
- **(d)** Left-handed helix

16. Which one of the following statements regarding silk fibroin is **incorrect**?
- **(a)** Fibroin forms extensive β-sheets
- **(b)** Fibroin has a high glycine and alanine content
- **(c)** It has a closely packed structure
- **(d)** The fibroin polypeptide forms a triple helix which gives it tensile strength

17. The structure in which two α-helices lie side by side in antiparallel directions in such a way that their side-chains intermesh is called what?
- **(a)** αα motif
- **(b)** βαβ loop
- **(c)** β turn
- **(d)** CD4 domain

18. Which of these proteins has a quaternary structure?
- **(a)** Carbonic anhydrase
- **(b)** Concanavalin A
- **(c)** Hemoglobin
- **(d)** Myoglobin

19. The tobacco mosaic virus capsid is made up of how many subunits?
- **(a)** 158
- **(b)** 240
- **(c)** 2130
- **(d)** 5200

20. The unfolding of a protein is called what?
- **(a)** Denaturation
- **(b)** Dialysis
- **(c)** Oxidation
- **(d)** Renaturation

21. What does the molten globule model for protein folding state?
- **(a)** Because the globule is molten it can change conformation rapidly
- **(b)** Collapse into a molten globule might automatically fold some of the polypeptide into its α-helices and β-sheets
- **(c)** The initial step in folding is the rapid collapse of the polypeptide into a compact structure
- **(d)** All of the above statements are part of the molten globule model

22. Hsp70 proteins are examples of what?
- **(a)** Chaperonins
- **(b)** Molecular chaperones
- **(c)** Molten globules
- **(d)** Motor proteins

23. The GroEL/GroES complex is a type of what?
- **(a)** Chaperonin
- **(b)** Hsp70 protein
- **(c)** Molten globule
- **(d)** Motor protein

24. Which one of the following is an example of a storage protein?
- **(a)** Dynein
- **(b)** Ferritin
- **(c)** Insulin
- **(d)** Keratin