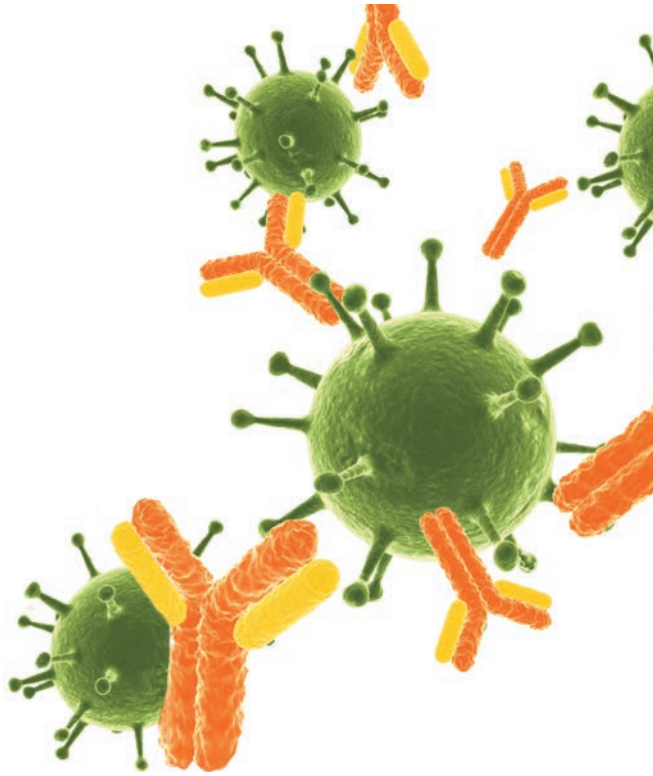


CHAPTER

24

Amino Acids and Proteins

Among the major classes of biomolecules, proteins arguably have the most diverse array of functions. As enzymes they serve as catalysts to affect chemical reactions; as antibodies they protect from disease; as molecules they form critical structures, including skin, hair, and nails; and as hormones they control many body functions, including metabolism, growth, and reproduction. As we shall see, not only do proteins come in all shapes and sizes, but each individual protein is also the product of an evolutionary process that has led to its specific properties and functions. By learning about the overall structure and function of proteins, chemists are now able to apply that knowledge to the development of some highly valuable, fully synthetic proteins of their own design. For instance, as we will see later in this chapter, lessons learned from the natural adaptability of antibodies generated by the immune system provided insights for the development of unnatural, synthetic analogs that can catalyze chemical reactions such as the Claisen rearrangement, aldol reactions, and the Diels–Alder reaction, as in the graphic shown to the right.

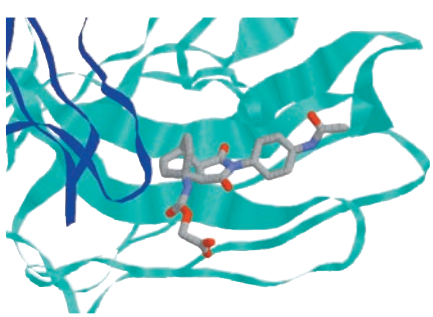


PHOTO CREDIT: Ingram Publishing/Getty Images, Inc. ILLUSTRATION CREDIT: PDB ID: 1A4K, http://www.pdb.org. Romesberg, F. E., Spiller, B., Schultz, P. G., Stevens, R. C. Immunological origins of binding and catalysis in a Diels–Alderase antibody. Science 279, pp. 1929–1933, 1998.

N THIS CHAPTER WE WILL CONSIDER:

- the structures and properties of amino acids that constitute proteins
- methods to determine the amino acid sequence of a given protein, as well as synthesize it
- the primary, secondary, tertiary, and quaternary structures of proteins
- selected examples of enzymes and their function

[WHY DO THESE TOPICS MATTER?] Not only will we show you how some novel catalytic antibodies work, but also at the end of this chapter we will show how chemists have gone beyond the standard amino acids found in nature to create proteins in diverse cells that include many new and fully synthetic amino acids.

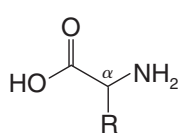
24.1 INTRODUCTION

The three groups of biological polymers are polysaccharides, proteins, and nucleic acids. We studied polysaccharides in Chapter 22 and saw that they function primarily as energy reserves, as biochemical labels on cell surfaces, and, in plants, as structural materials. When we study nucleic acids in Chapter 25, we shall find that they serve two major purposes: storage and transmission of information. Of the three groups of biopolymers, proteins have the most diverse functions. As enzymes and hormones, proteins catalyze and regulate the reactions that occur in the body; as muscles and tendons they provide the body with the means for movement; as skin and hair they give it an outer covering; as hemoglobin molecules they transfer all-important oxygen to its most remote corners; as antibodies they provide it with a means of protection against disease; and in combination with other substances in bone they provide it with structural support.

Given such diversity of functions, we should not be surprised to find that proteins come in all sizes and shapes. By the standard of most of the molecules we have studied, even small proteins have very high molecular weights. Lysozyme, an enzyme, is a relatively small protein and yet its molecular weight is 14,600. The molecular weights of most proteins are much larger. Their shapes cover a range from the globular proteins such as lysozyme and hemoglobin to the helical coils of α-keratin (hair, nails, and wool) and the pleated sheets of silk fibroin.

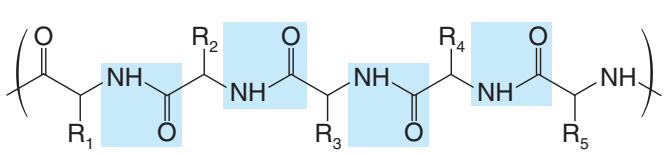
And yet, in spite of such diversity of size, shape, and function, all proteins have common features that allow us to deduce their structures and understand their properties. Later in this chapter we shall see how this is done.

- Proteins are **polyamides**, and their monomeric units are composed of about 20 different α-amino acids:



An α-amino acid

R is a side chain at the α carbon that determines the identity of the amino acid (Table 24.1).



A portion of a protein molecule

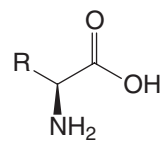
Amide (peptide) linkages are shaded.
R_1 to R_5 may be any of the possible side chains.

- The exact sequence of the different α-amino acids along the protein chain is called the **primary structure** of the protein.

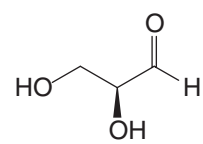
A protein's primary structure, as its name suggests, is of fundamental importance. For the protein to carry out its particular function, the primary structure must be correct. We shall see later that when the primary structure is correct, the protein's polyamide chain folds in particular ways to give it the shape it needs for its particular task.

- Folding of the polyamide chain gives rise to higher levels of complexity called the **secondary** and **tertiary structures** of the protein.
- **Quaternary structure** results when a protein contains an aggregate of more than one polyamide chain.
- Hydrolysis of proteins with acid or base yields a mixture of amino acids.

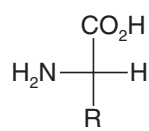
Although hydrolysis of naturally occurring proteins may yield as many as 22 different amino acids, the amino acids have an important structural feature in common: with the exception of glycine (whose molecules are achiral), almost all naturally occurring amino acids have the L configuration at the α carbon.* That is, they have the same relative configuration as L-glyceraldehyde:

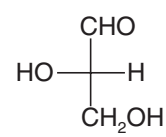


An L-α-amino acid [usually an (*S*)-α-amino acid]



L-Glyceraldehyde [(*S*)-glyceraldehyde]





Fischer projections for an L-α-amino acid and L-glyceraldehyde

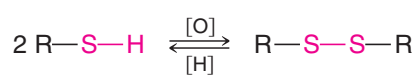
24.2 AMINO ACIDS

24.2A Structures and Names

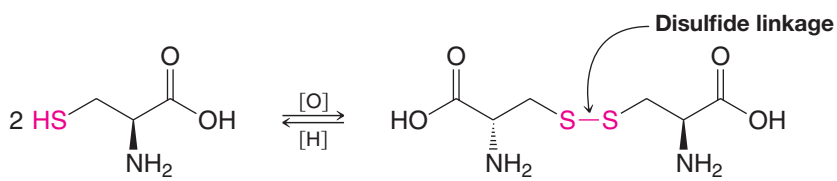
- The 22 α-amino acids that can be obtained from proteins can be subdivided into three different groups on the basis of the structures of their side chains, R. These are given in Table 24.1.

Only 20 of the 22 α-amino acids in Table 24.1 are actually used by cells when they synthesize proteins. Two amino acids are synthesized after the polyamide chain is intact. Hydroxyproline (present mainly in collagen) is synthesized by oxidation of proline, and cystine (present in most proteins) is synthesized from cysteine.

The conversion of cysteine to cystine requires additional comment. The —SH group of cysteine makes cysteine a *thiol*. One property of thiols is that they can be converted to disulfides by mild oxidizing agents. This conversion, moreover, can be reversed by mild reducing agents:



Thiol **Disulfide**

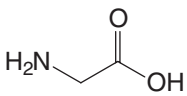
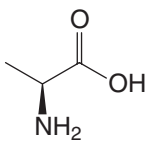
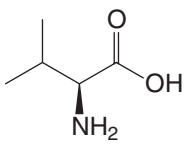
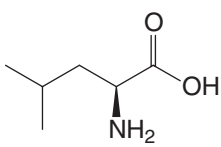
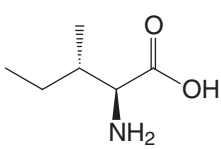
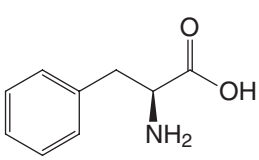
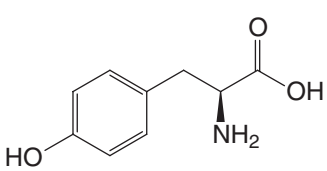
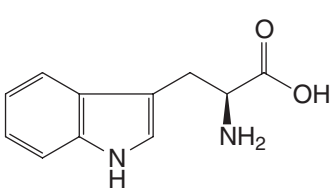
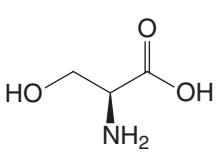
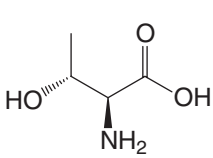


Cysteine **Cystine**

We shall see later how the **disulfide linkage** between cysteine units in a protein chain contributes to the overall structure and shape of the protein.

*Some D-amino acids have been obtained from the material comprising the cell walls of bacteria and by hydrolysis of certain antibiotics.

TABLE 24.1 L-AMINO ACIDS FOUND IN PROTEINS

Structure	Name	Abbreviations[a]	pK_{a_1} α-CO_2H	pK_{a_2} α-NH_3^+	pK_{a_3} R group	p*I*
Neutral Amino Acids						
	Glycine	G or Gly	2.3	9.6		6.0
	Alanine	A or Ala	2.3	9.7		6.0
	Valine[b]	V or Val	2.3	9.6		6.0
	Leucine[b]	L or Leu	2.4	9.6		6.0
	Isoleucine[b]	I or Ile	2.4	9.7		6.1
	Phenylalanine[b]	F or Phe	1.8	9.1		5.5
	Tyrosine	Y or Tyr	2.2	9.1	10.1	5.7
	Tryptophan[b]	W or Trp	2.4	9.4		5.9
	Serine	S or Ser	2.2	9.2		5.7
	Threonine[b]	T or Thr	2.6	10.4		6.5

(continues on next page)

TABLE 24.1 CONTINUED

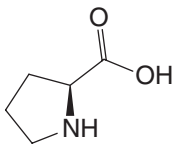
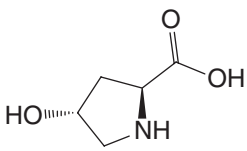
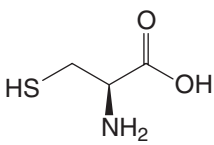
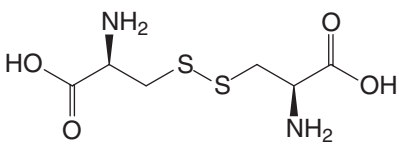
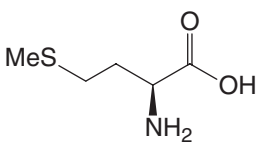
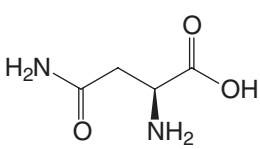
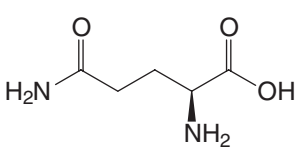
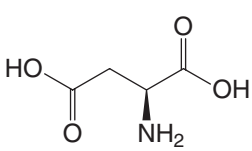
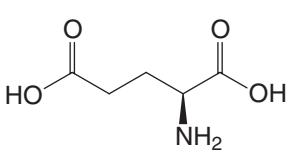
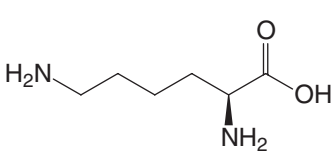
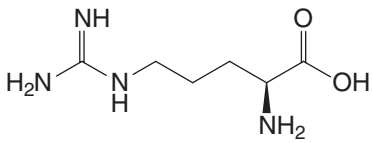
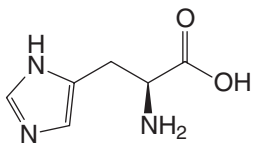
Structure	Name	Abbreviations[a]	pK_{a_1} α-CO_2H	pK_{a_2} α-NH_3^+	pK_{a_3} R group	pI
	Proline	P or Pro	2.0	10.6		6.3
	4-Hydroxyproline (cis and trans)	O or Hyp	1.9	9.7		6.3
	Cysteine	C or Cys	1.7	10.8	8.3	5.0
	Cystine	Cys-Cys	1.6 2.3	7.9 9.9		5.1
	Methionine[b]	M or Met	2.3	9.2		5.8
	Asparagine	N or Asn	2.0	8.8		5.4
	Glutamine	Q or Gln	2.2	9.1		5.7
Side Chains Containing an Acidic (Carboxyl) Group						
	Aspartic acid	D or Asp	2.1	9.8	3.9	3.0
	Glutamic acid	E or Glu	2.2	9.7	4.3	3.2
Side Chains Containing a Basic Group						
	Lysine[b]	K or Lys	2.2	9.0	10.5[c]	9.8

TABLE 24.1 CONTINUED

Structure	Name	Abbreviations[a]	pK_{a_1} α-CO_2H	pK_{a_2} α-NH_3^+	pK_{a_3} R group	p*I*
	Arginine	R or Arg	2.2	9.0	12.5[c]	10.8
	Histidine	H or His	1.8	9.2	6.0[c]	7.6

[a]Single-letter abbreviations are now the most commonly used form in current biochemical literature.

[b]An essential amino acid.

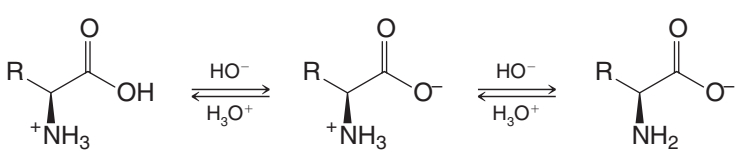
[c]pK_a is of protonated amine of R group.

24.2B Essential Amino Acids

Amino acids can be synthesized by all living organisms, plants and animals. Many higher animals, however, are deficient in their ability to synthesize all of the amino acids they need for their proteins. Thus, these higher animals require certain amino acids as a part of their diet. For adult humans there are eight **essential amino acids**; these are identified in Table 24.1 by a footnote.

24.2C Amino Acids as Dipolar Ions

- Amino acids contain both a basic group ($—NH_2$) and an acidic group ($—CO_2H$).
- In the dry solid state, amino acids exist as **dipolar ions**, a form in which the carboxyl group is present as a carboxylate ion, $—CO_2^-$, and the amino group is present as an aminium ion, $—NH_3^+$. (Dipolar ions are also called **zwitterions**.)
- In aqueous solution, an equilibrium exists between the dipolar ion and the anionic and cationic forms of an amino acid.



Cationic form (predominant in strongly acidic solutions, e.g., at pH 0) | **Dipolar ion** | **Anionic form (predominant in strongly basic solutions, e.g., at pH 14)**

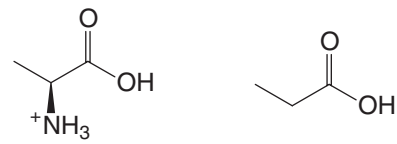
The predominant form of the amino acid present in a solution depends on the pH of the solution and on the nature of the amino acid. In strongly acidic solutions all amino acids are present primarily as cations; in strongly basic solutions they are present as anions.

- The **isoelectric point (p*I*)** is the pH at which the concentration of the dipolar ion is at its maximum and the concentrations of the anions and cations are equal.

Each amino acid has a particular isoelectric point. These are given in Table 24.1. Proteins have isoelectric points as well. As we shall see later (Sections 24.13 and 24.14), this property of proteins is important for their separation and identification.

Let us consider first an amino acid with a side chain that contains neither acidic nor basic groups—an amino acid, for example, such as alanine.

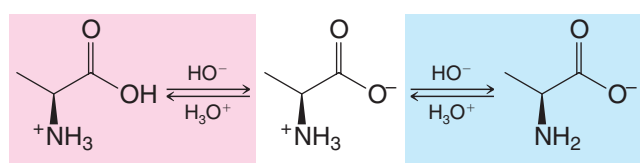
If alanine is dissolved in a strongly acidic solution (e.g., pH 0), it is present in mainly a net cationic form. In this state the amino group is protonated (bears a formal +1 charge) and the carboxylic acid group is neutral (has no formal charge). As is typical of α-amino acids, the pK_a for the carboxylic acid hydrogen of alanine is considerably lower (2.3) than the pK_a of an ordinary carboxylic acid (e.g., propanoic acid, pK_a 4.89):



Cationic form of alanine $pK_{a_1} = 2.3$ | **Propanoic acid** $pK_a = 4.89$

The reason for this enhanced acidity of the carboxyl group in an α-amino acid is the inductive effect of the neighboring aminium cation, which helps to stabilize the carboxylate anion formed when it loses a proton. Loss of a proton from the carboxyl group in a cationic α-amino acid leaves the molecule electrically neutral (in the form of a dipolar ion). This equilibrium is shown in the red-shaded portion of the equation below.

The protonated amino group of an α-amino acid is also acidic, but less so than the carboxylic acid group. The pK_a of the aminium group in alanine is 9.7. The equilibrium for loss of an aminium proton is shown in the blue-shaded portion of the equation below. The carboxylic acid proton is always lost before a proton from the aminium group in an α-amino acid.



Cationic form ($pK_{a_1} = 2.3$) | **Dipolar ion** ($pK_{a_2} = 9.7$) | **Anionic form**

The state of an α-amino acid at any given pH is governed by a combination of two equilibria, as shown in the above equation for alanine. The isoelectric point (pI) of an amino acid such as alanine is the average of pK_{a_1} and pK_{a_2}:

$$pI = \tfrac{1}{2}(2.3 + 9.7) = 6.0 \qquad \text{(isoelectric point of alanine)}$$

When a base is added to a solution of the net cationic form of alanine (initially at pH 0, for example), the first proton removed is the carboxylic acid proton, as we have said. In the case of alanine, when a pH of 2.3 is reached, the carboxylic acid proton will have been removed from half of the molecules. This pH represents the pK_a of the alanine carboxylic acid proton, as can be demonstrated using the **Henderson–Hasselbalch equation**.

- The **Henderson–Hasselbalch equation** shows that for an acid (HA) and its conjugate base (A^-) when [HA] = [A^-], then pH = pK_a.

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

Henderson–Hasselbalch equation

Therefore, when the acid is half neutralized,

$$[HA] = [A^-],\ \log \frac{[HA]}{[A^-]} = 0, \text{ and thus } pH = pK_a$$

As more base is added to this solution, alanine reaches its isoelectric point (p*I*), the pH at which all of alanine's carboxylic acid protons have been removed but not its aminium protons. The molecules are therefore electrically neutral (in their dipolar ion or zwitter ionic form) because the carboxylate group carries a −1 charge and the aminium group a +1 charge. The p*I* for alanine is 6.0.

Now, as we continue to add the base, protons from the aminium ions will begin to be removed, until at pH 9.7 half of the aminium groups will have lost a proton. This pH represents the pK_a of the aminium group. Finally, as more base is added, the remaining aminium protons will be lost until all of the alanine molecules have lost their aminium protons. At this point (e.g., pH 14) the molecules carry a net anionic charge from their carboxylate group. The amino groups are now electrically neutral.

Figure 24.1 shows a titration curve for these equilibria. The graph represents the change in pH as a function of the number of molar equivalents of base. Because alanine has two protons to lose in its net cationic form, when one molar equivalent of base has been added, the molecules will have each lost one proton and they will be electrically neutral (the dipolar ion or zwitterionic form).

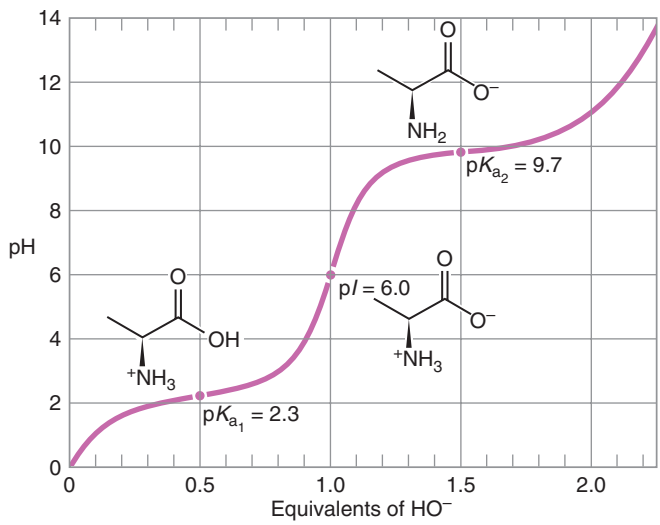
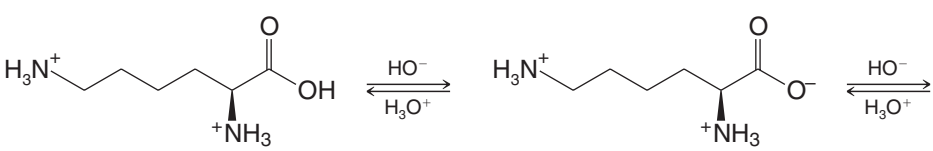
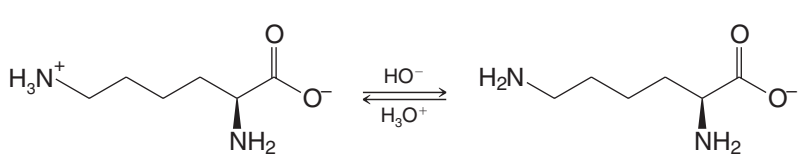


FIGURE 24.1 A titration curve for alanine.

If an amino acid contains a side chain that has an acidic or basic group, the equilibria become more complex. Consider lysine, for example, an amino acid that has an additional —NH_2 group on its ε carbon. In strongly acidic solution, lysine is present as a dication because both amino groups are protonated. The first proton to be lost as the pH is raised is a proton of the carboxyl group (pK_{a_1} = 2.2), the next is from the α-aminium group (pK_{a_2} = 9.0), and the last is from the ε-aminium group (pK_{a_3} = 10.5):



Dicationic form of lysine (pK_{a_1} = 2.2) — **Monocationic form (pK_{a_2} = 9.0)**



Dipolar ion (pK_{a_3} = 10.5) — **Anionic form**

The isoelectric point of lysine is the average of pK_{a_2} (the monocation) and pK_{a_3} (the dipolar ion).

$$pI = \tfrac{1}{2}(9.01 + 0.5) = 9.8 \quad \text{(isoelectric point of lysine)}$$

PRACTICE PROBLEM 24.1 What form of glutamic acid would you expect to predominate in **(a)** strongly acidic solution, **(b)** strongly basic solution, and **(c)** at its isoelectric point ($pI = 3.2$)? **(d)** The isoelectric point of glutamine ($pI = 5.7$) is considerably higher than that of glutamic acid. Explain.

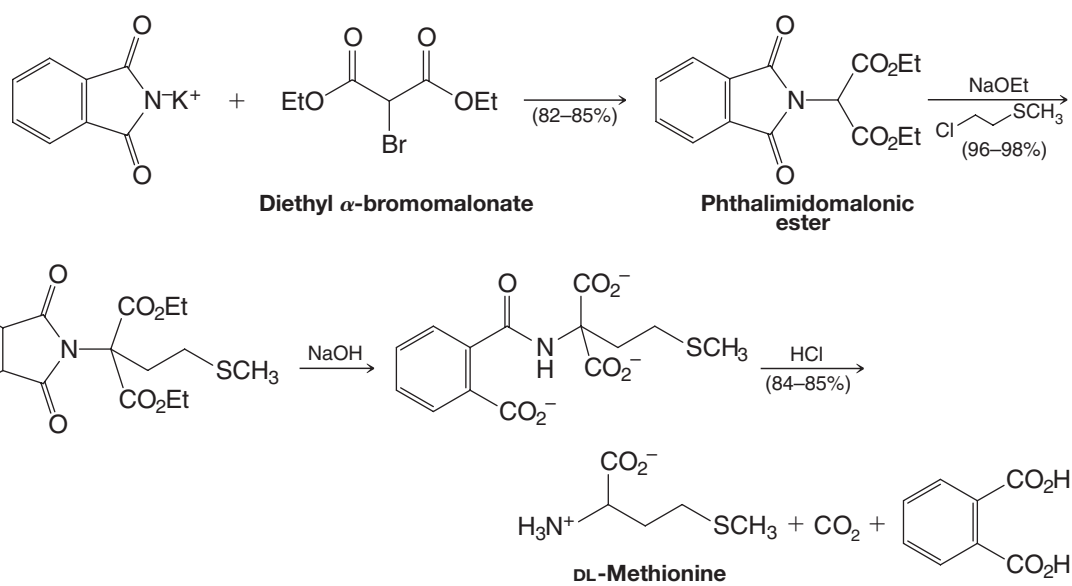
PRACTICE PROBLEM 24.2 The guanidino group —NH—C(=NH)—NH_2 of arginine is one of the most strongly basic of all organic groups. Explain.

24.3 SYNTHESIS OF α-AMINO ACIDS

A variety of methods have been developed for the synthesis of α-amino acids. Here we describe two methods that are based on reactions we have studied before. In "The Chemistry of . . . Asymmetric Syntheses of Amino Acids" (*WileyPLUS*) we show methods to prepare α-amino acids in optically active form. Asymmetric synthesis is an important goal in α-amino acid synthesis due to the biological activity of the natural enantiomeric forms of α-amino acids, and due to the commercial relevance of products made by these routes.

24.3A From Potassium Phthalimide

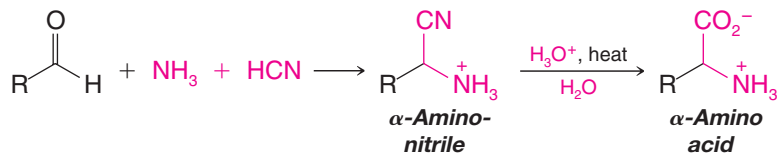
This method, a modification of the Gabriel synthesis of amines (Section 20.4A), uses potassium phthalimide and diethyl α-bromomalonate to prepare an *imido* malonic ester. The example shown is a synthesis of methionine:



PRACTICE PROBLEM 24.3 Starting with diethyl α-bromomalonate and potassium phthalimide and using any other necessary reagents, show how you might synthesize **(a)** DL-leucine, **(b)** DL-alanine, and **(c)** DL-phenylalanine.

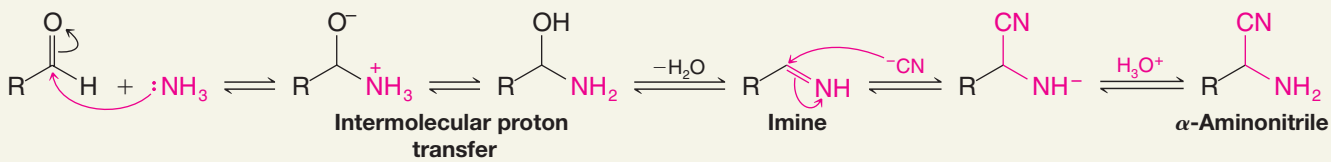
24.3B The Strecker Synthesis

Treating an aldehyde with ammonia and hydrogen cyanide produces an α-aminonitrile. Hydrolysis of the nitrile group (Section 17.3) of the α-aminonitrile converts the latter to an α-amino acid. This synthesis is called the Strecker synthesis:



The first step of this synthesis probably involves the initial formation of an imine from the aldehyde and ammonia followed by the addition of hydrogen cyanide.

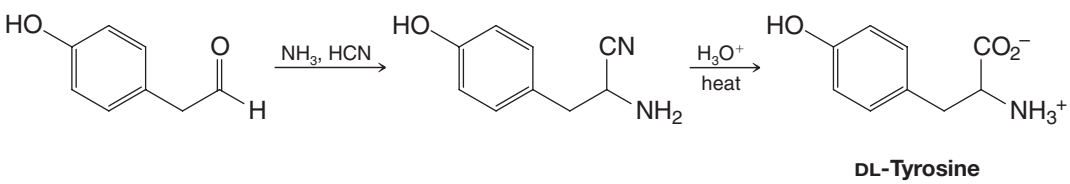
A MECHANISM FOR THE REACTION — Formation of an α-Aminonitrile during the Strecker Synthesis



SOLVED PROBLEM 24.1

Outline a Strecker synthesis of DL-tyrosine.

ANSWER:



DL-Tyrosine

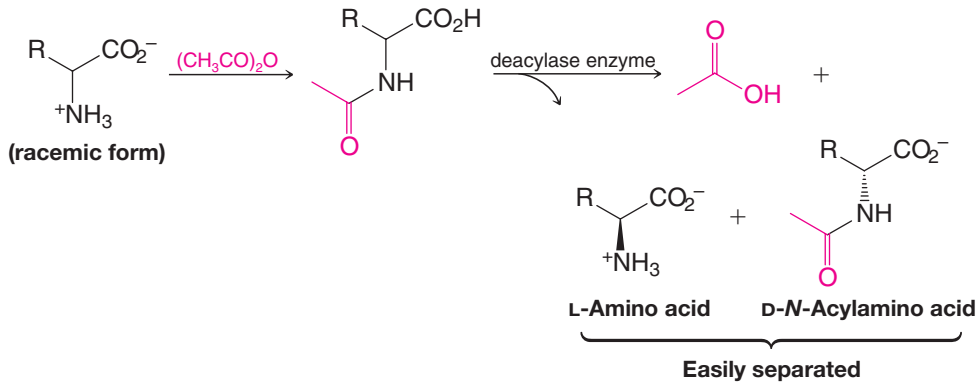
PRACTICE PROBLEM 24.4

(a) Outline a Strecker synthesis of DL-phenylalanine. **(b)** DL-Methionine can also be synthesized by a Strecker synthesis. The required starting aldehyde can be prepared from acrolein ($CH_2{=}CHCHO$) and methanethiol (CH_3SH). Outline all steps in this synthesis of DL-methionine.

24.3C Resolution of DL-Amino Acids

With the exception of glycine, which has no chirality center, the amino acids that are produced by the methods we have outlined are all produced as racemic forms. To obtain the naturally occurring L-amino acid, we must, of course, resolve the racemic form. This can be done in a variety of ways, including the methods outlined in Section 20.3F.

One especially interesting method for resolving amino acids is based on the use of enzymes called *deacylases*. These enzymes catalyze the hydrolysis of *N-acylamino acids* in living organisms. Since the active site of the enzyme is chiral, it hydrolyzes only *N*-acylamino acids of the L configuration. When it is exposed to a racemic mixture of *N*-acylamino acids, only the derivative of the L-amino acid is affected and the products, as a result, are separated easily:

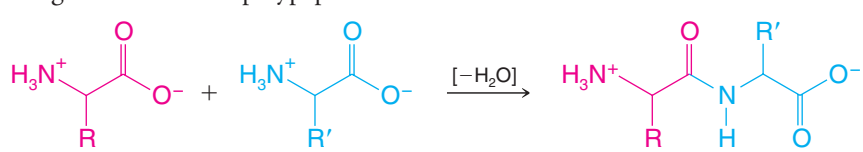


24.4 POLYPEPTIDES AND PROTEINS

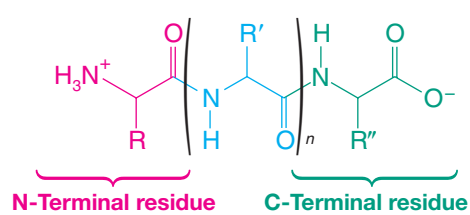
Amino acids are polymerized in living systems by enzymes that form amide linkages from the amino group of one amino acid to the carboxyl group of another.

- A molecule formed by joining amino acids together is called a **peptide**, and the amide linkages in them are called **peptide bonds** or **peptide linkages**. Each amino acid in the peptide is called an **amino acid residue**.

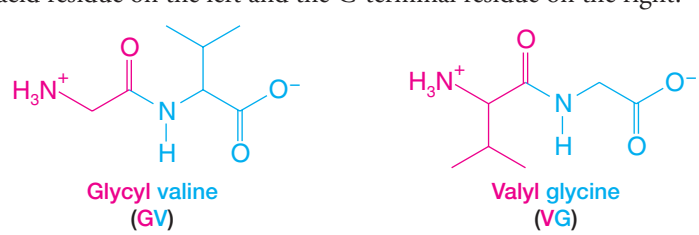
Peptides that contain 2, 3, a few (3–10), or many amino acids are called **dipeptides**, **tripeptides**, **oligopeptides**, and **polypeptides**, respectively. **Proteins** are polypeptides consisting of one or more polypeptide chains.



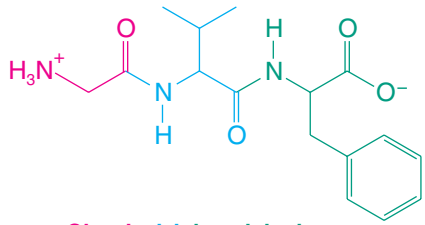
Polypeptides are **linear polymers**. One end of a polypeptide chain terminates in an amino acid residue that has a free $-NH_3^+$ group; the other terminates in an amino acid residue with a free $-CO_2^-$ group. These two groups are called the **N-terminal** and the **C-terminal residues**, respectively:



- By convention, we write peptide and protein structures with the N-terminal amino acid residue on the left and the C-terminal residue on the right:



The tripeptide glycylvalylphenylalanine has the following structural formula:



Glycylvalylphenylalanine
(GVF)

It becomes a significant task to write a full structural formula for a polypeptide chain that contains any more than a few amino acid residues. In this situation, use of the one-letter abbreviations (Table 24.1) is the norm for showing the sequence of amino acids. Very short peptide sequences are sometimes still represented with the three-letter abbreviations (Table 24.1).

24.4A Hydrolysis

When a protein or polypeptide is refluxed with 6 M hydrochloric acid for 24 h, hydrolysis of all the amide linkages usually takes place, liberating its constitutent amino acids as a mixture. Chromatographic separation and quantitative analysis of the resulting mixture can then be used to determine which amino acids composed the intact polypeptide and their relative amounts.

One chromatographic method for separation of a mixture of amino acids is based on the use of *cation-exchange resins* (Fig. 24.2), which are insoluble polymers containing sulfonate groups. If an acidic solution containing a mixture of amino acids is passed through a column packed with a cation-exchange resin, the amino acids will be adsorbed by the resin because of attractive forces between the negatively charged sulfonate groups and the positively charged amino acids. The strength of the adsorption varies with the basicity of the individual amino acids; those that are most basic are held most strongly. If the column is then washed with a buffered solution at a given pH, the individual amino acids move down the column at different rates and ultimately become separated. In an automated version of this analysis developed at Rockefeller University in 1950, the eluate is allowed to mix with **ninhydrin**, a reagent that reacts with most amino acids to give a derivative with an intense purple color (λ_{max} 570 nm). The amino acid analyzer is designed so that it can measure the absorbance of the eluate (at 570 nm) continuously and record this absorbance as a function of the volume of the effluent.

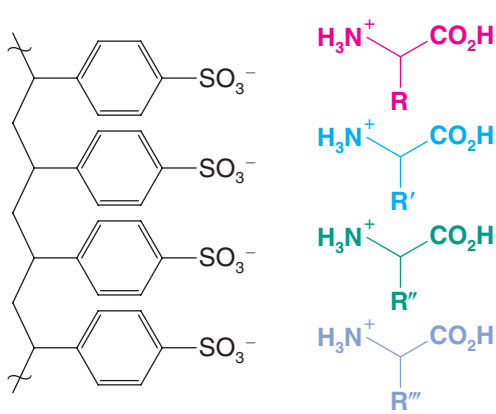


FIGURE 24.2 A section of a cation-exchange resin with adsorbed amino acids.

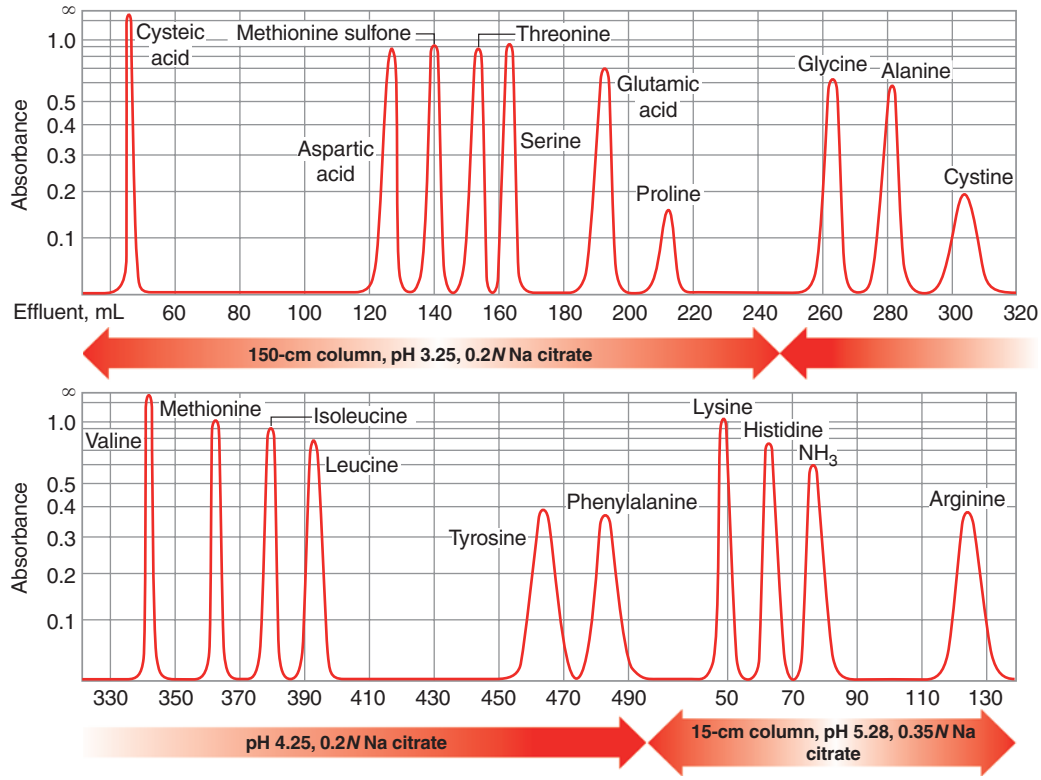
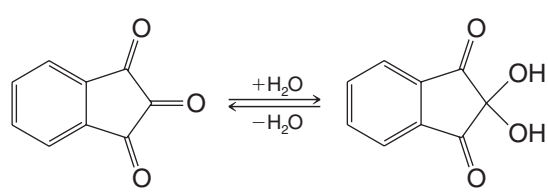


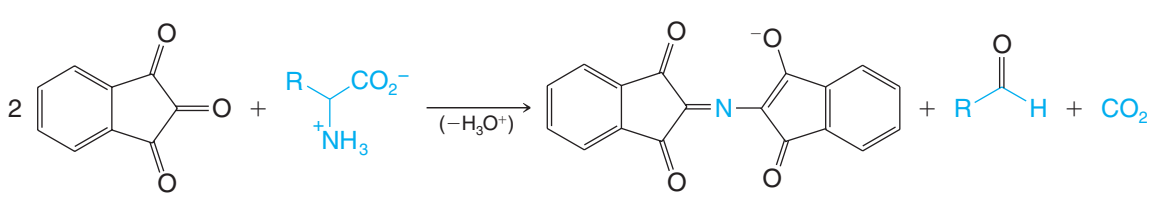
FIGURE 24.3 Typical result given by an automatic amino acid analyzer. (Adapted with permission from Spackman, D. H., Stein, W. H., and Moore, S., *Analytical Chemistry, 30(7)*, pp. 1190–1206, Figure 2, 1958. Copyright 1958 American Chemical Society.)

A typical graph obtained from an automatic amino acid analyzer is shown in Fig. 24.3. When the procedure is standardized, the positions of the peaks are characteristic of the individual amino acids, and the areas under the peaks correspond to their relative amounts.

Ninhydrin is the hydrate of indane-1,2,3-trione. With the exception of proline and hydroxyproline, all of the α-amino acids found in proteins react with ninhydrin to give the same intensely colored purple anion (λ_{max} 570 nm). We shall not go into the mechanism here, but notice that the only portion of the anion that is derived from the α-amino acid is the nitrogen:



Indane-1,2,3-trione **Ninhydrin**



Purple anion

Proline and hydroxyproline do not react with ninhydrin in the same way because their α-amino groups are secondary amines and part of a five-membered ring.

Analysis of amino acid mixtures can also be done very easily using high-performance liquid chromatography (HPLC), and this is now the most common method. A cation-exchange resin is used for the column packing in some HPLC analyses

(see Section 24.14), while other analyses require hydrophobic (reversed-phase) column materials. Identification of amino acids separated by HPLC can be done by comparison with retention times of standard samples. Instruments that combine HPLC with mass spectrometry make direct identification possible (see Section 24.5E).

24.5 PRIMARY STRUCTURE OF POLYPEPTIDES AND PROTEINS

The sequence of amino acid residues in a polypeptide or protein is called its **primary structure**. A simple peptide composed of three amino acids (a tripeptide) can have 6 different amino acid sequences; a tetrapeptide can have as many as 24 different sequences. For a protein composed of 20 different amino acids in a single chain of 100 residues, there are $2^{100} = 1.27 \times 10^{130}$ possible peptide sequences, a number much greater than the number of atoms estimated to be in the universe (9×10^{78})! Clearly, one of the most important things to determine about a protein is the sequence of its amino acids. Fortunately, there are a variety of methods available to determine the sequence of amino acids in a polypeptide. We shall begin with **terminal residue analysis** techniques used to identify the N- and C-terminal amino acids.

24.5A Edman Degradation

The most widely used procedure for identifying the N-terminal amino acid in a peptide is the **Edman degradation** method (developed by Pehr Edman of the University of Lund, Sweden). Used repetitively, the Edman degradation method can be used to sequence peptides up to about 60 residues in length. The process works so well that machines called amino acid sequencers have been developed to carry out the Edman degradation process in automated cycles.

The chemistry of the Edman degradation is based on a labeling reaction between the N-terminal amino group and phenyl isothiocyanate, $C_6H_5—N{=}C{=}S$. Phenyl isothiocyanate reacts with the N-terminal amino group to form a phenylthiocarbamyl derivative, which is then cleaved from the peptide chain by acid. The result is an unstable anilinothioazolinone (ATZ), which rearranges to a stable phenylthiohydantoin (PTH) derivative of the amino acid. In the automated process, the PTH derivative is introduced directly to a high-performance liquid chromatograph and identified by comparison of its retention time with known amino acid PTH derivatives (Fig. 24.4). The cycle is then repeated for the next N-terminal amino acid. Automated peptide sequence analyzers can

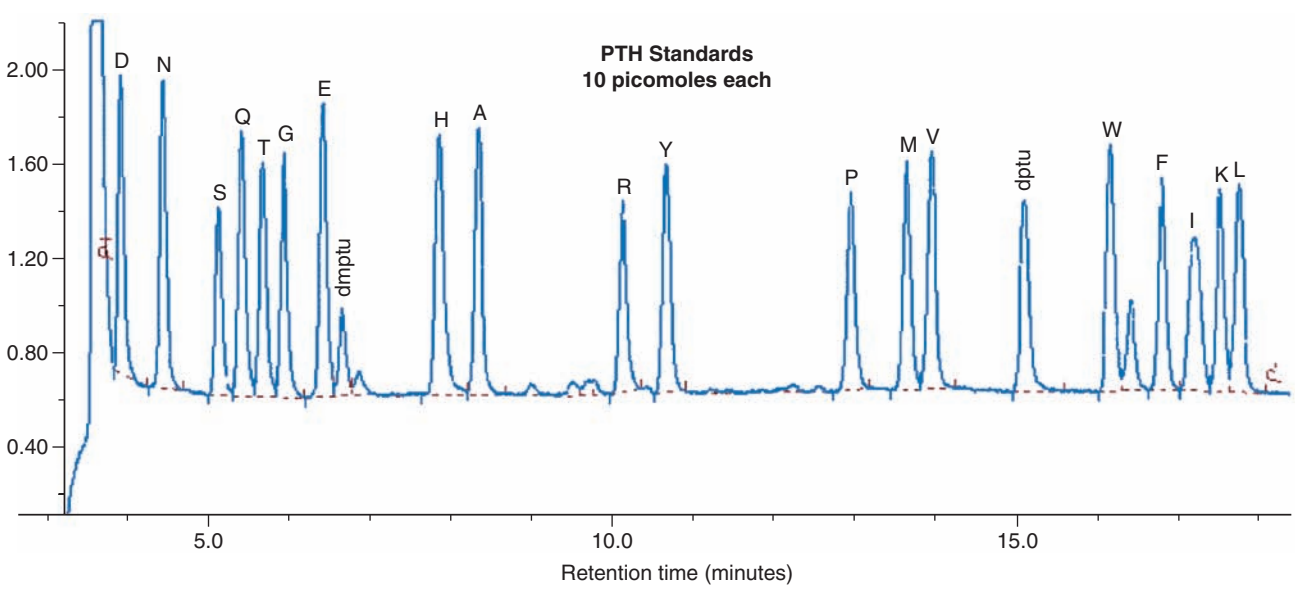
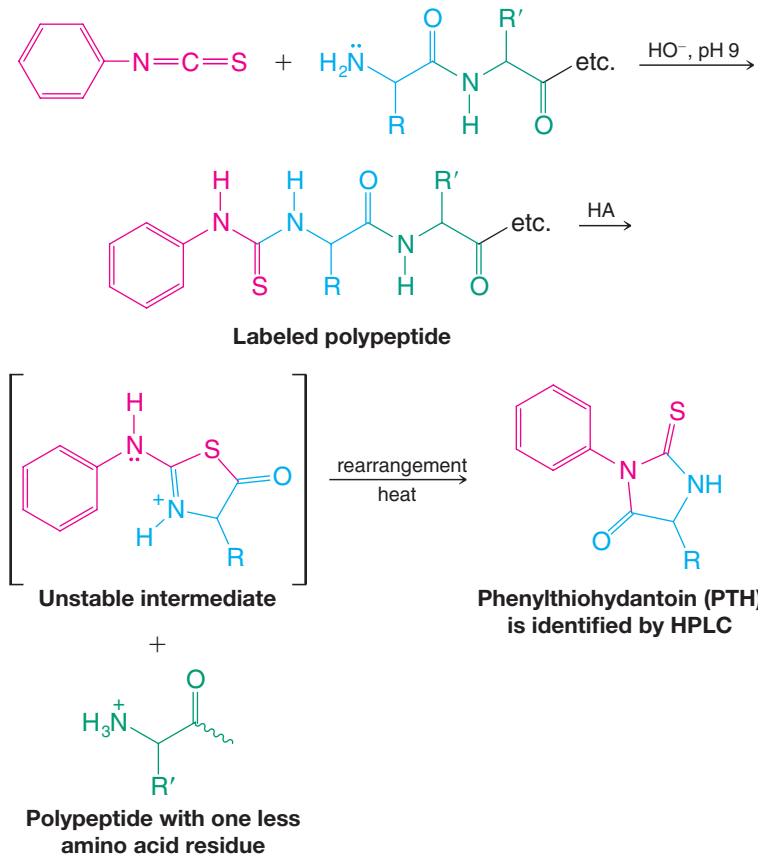


FIGURE 24.4 PTH amino acid standards run on a Procise instrument; see Table 24.1 for amino acid abbreviations. Peaks marked dmptu (dimethylphenylthiourea) and dptu (diphenylthiourea) represent side-reaction products of the Edman degradation. (Copyright © *2012 Life Technologies Corporation. Used under permission.*)

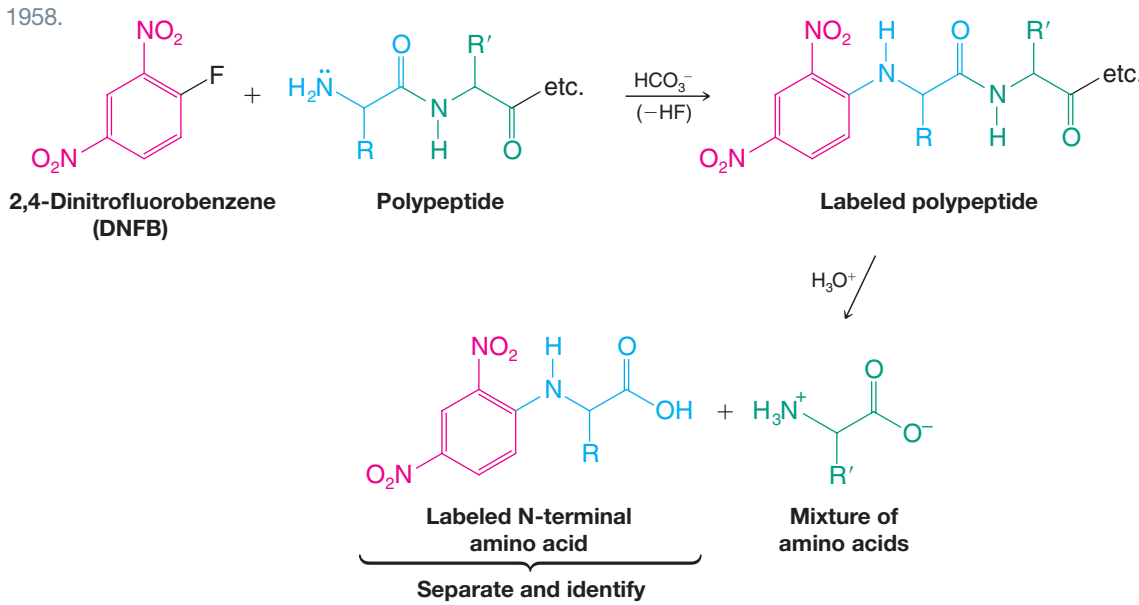
perform a single iteration of the Edman degradation in approximately 30 min using only picomole amounts of the polypeptide sample.

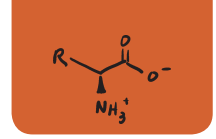


This method was introduced by Frederick Sanger of Cambridge University in 1945. Sanger made extensive use of this procedure in his determination of the amino acid sequence of insulin and won the Nobel Prize in Chemistry for the work in 1958.

24.5B Sanger N-Terminal Analysis

Another method for sequence analysis is the **Sanger N-terminal analysis**, based on the use of 2,4-dinitrofluorobenzene (DNFB). When a polypeptide is treated with DNFB in mildly basic solution, a nucleophilic aromatic substitution reaction (S_NAr, Section 21.11A) takes place involving the free amino group of the N-terminal residue. Subsequent hydrolysis of the polypeptide gives a mixture of amino acids in which the N-terminal amino acid is labeled with a 2,4-dinitrophenyl group. After separating this amino acid from the mixture, it can be identified by comparison with known standards.





2,4-Dinitrofluorobenzene will react with any free amino group in a polypeptide, including the ε-amino group of lysine, and this fact complicates Sanger analyses. Only the N-terminal amino acid residue of a peptide will bear the 2,4-dinitrophenyl group at its α-amino group, however. Nevertheless, the Edman method of N-terminal analysis is much more widely used.

PRACTICE PROBLEM 24.5

The electron-withdrawing property of the 2,4-dinitrophenyl group makes separation of the labeled amino acid very easy. Suggest how this is done.

24.5C C-Terminal Analysis

C-Terminal residues can be identified through the use of digestive enzymes called *carboxypeptidases.* These enzymes specifically catalyze the hydrolysis of the amide bond of the amino acid residue containing a free $-CO_2H$ group, liberating it as a free amino acid. A carboxypeptidase, however, will continue to attack the polypeptide chain that remains, successively lopping off C-terminal residues. As a consequence, it is necessary to follow the amino acids released as a function of time. The procedure can be applied to only a limited amino acid sequence for, at best, after a time the situation becomes too confused to sort out.

PRACTICE PROBLEM 24.6

(a) Write a reaction showing how 2,4-dinitrofluorobenzene could be used to identify the N-terminal amino acid of VAG. **(b)** What products would you expect (after hydrolysis) when VKG is treated with 2,4-dinitrofluorobenzene?

PRACTICE PROBLEM 24.7

Write the reactions involved in a sequential Edman degradation of MIR.

24.5D Complete Sequence Analysis

Sequential analysis using the Edman degradation or other methods becomes impractical with large proteins and polypeptides. Fortunately, there are techniques to cleave peptides into fragments that are of manageable size. **Partial hydrolysis** with dilute acid, for example, generates a family of peptides cleaved in random locations and with varying lengths. Sequencing these cleavage peptides and looking for points of overlap allows the sequence of the entire peptide to be pieced together.

Consider a simple example: we are given a pentapeptide known to contain valine (two residues), leucine (one residue), histidine (one residue), and phenylalanine (one residue), as determined by hydrolysis and automatic amino acid analysis. With this information we can write the "molecular formula" of the protein in the following way, using commas to indicate that the sequence is unknown:

2V, L, H, F

Then, let us assume that by using DNFB and carboxypeptidase we discover that valine and leucine are the N- and C-terminal residues, respectively. So far we know the following:

V (V, H, F) L

But the sequence of the three nonterminal amino acids is still unknown.

We then subject the pentapeptide to partial acid hydrolysis and obtain the following dipeptides. (We also get individual amino acids and larger pieces, i.e., tripeptides and tetrapeptides.)

VH + HV + VF + FL

The points of overlap of the dipeptides (i.e., H, V, and F) tell us that the original pentapeptide must have been the following:

VHVFL

Site-specific cleavage of peptide bonds is possible with enzymes and specialized reagents as well, and these methods are now more widely used than partial hydrolysis. For example, the enzyme trypsin preferentially catalyzes hydrolysis of peptide bonds on the C-terminal side of arginine and lysine. Chemical cleavage at specific sites can be done with cyanogen bromide (CNBr), which cleaves peptide bonds on the C-terminal side of methionine residues. Using these site-selective cleavage methods on separate samples of a given polypeptide results in fragments that have overlapping sequences. After sequencing the individual fragments, aligning them with each other on the basis of their overlapping sections results in a sequence for the intact protein.

24.5E Peptide Sequencing Using Mass Spectrometry and Sequence Databases

Other methods for determining the sequence of a polypeptide include mass spectrometry and comparison of partial peptide sequences with databases of known complete sequences.

Ladder Sequencing Mass spectrometry is especially powerful because sophisticated techniques allow mass analysis of proteins with very high precision. In one mass spectrometric method, called "ladder sequencing," an enzymatic digest is prepared that yields a mixture of peptide fragments that each differ in length by one amino acid residue (e.g., by use of carboxypeptidase). The digest is a family of peptides where each one is the result of cleavage of one successive residue from the chain. Mass spectrometric analysis of this mixture yields a family of peaks corresponding to the molecular weight of each peptide. Each peak in the spectrum differs from the next by the molecular weight of the amino acid that is the difference in their structures. With these data, one can ascend the ladder of peaks from the lowest weight fragment to the highest (or vice versa), "reading" the sequence of the peptide from the difference in mass between each peak. The difference in mass between each peptide fragment and the next represents the amino acid in that spot along the sequence, and hence an entire sequence can be read from the ladder of fragment masses. This technique has also been applied to the sequencing of oligonucleotides.

Tandem Mass Spectrometry (MS/MS) Random cleavage of a peptide, similar to that from partial hydrolysis with acid, can also be accomplished with mass spectrometry. An intact protein introduced into a mass spectrometer can be cleaved into smaller fragments by collision with gas molecules deliberately leaked into the mass spectrometer vacuum chamber (a technique called collision-induced dissociation, CID). These peptide fragments can be individually selected for analysis using a technique called tandem mass spectrometry (MS/MS). The mass spectra of these random fragments can be compared with mass spectra databases to determine the protein sequence.

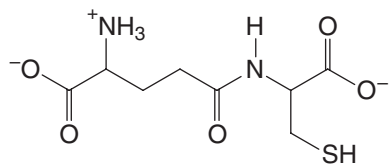
Partial Hydrolysis and Sequence Comparison In some cases it is also possible to determine the sequence of an unknown polypeptide by sequencing just a few of its amino acids and comparing this partial sequence with the database of known sequences for complete polypeptides or proteins. This procedure works if the unknown peptide turns out to be one that has been studied previously. (Studies of the expression of known proteins is one dimension of the field of proteomics, Section 24.14.) Due to the many sequence permutations that are theoretically possible and the uniqueness of a given protein's structure, a sequence of just 10–25 peptide residues is usually sufficient to generate data that match only one or a small number of known polypeptides. The partial sequence can be determined by the Edman method or by mass spectrometry. For example, the enzyme lysozyme with 129 amino acid residues (see Section 24.10) can be identified based on the sequence of just its first 15 amino acid residues. Structure determination based on comparison of sequences with computerized databases is part of the burgeoning field of bioinformatics.

An analogous approach using databases is to infer the *DNA sequence* that codes for a partial peptide sequence and compare this DNA sequence with the database of known

DNA sequences. If a satisfactory match is found, the remaining sequence of the polypeptide can be read from the DNA sequence using the genetic code (see Section 25.5). In addition, the inferred oligonucleotide sequence for the partial peptide can be synthesized chemically (see Section 25.7) and used as a probe to find the gene that codes for the protein. This technique is part of molecular biological methods used to clone and express large quantities of a protein of interest.

PRACTICE PROBLEM 24.8

Glutathione is a tripeptide found in most living cells. Partial acid-catalyzed hydrolysis of glutathione yields two dipeptides, CG and one composed of E and C. When this second dipeptide was treated with DNFB, acid hydrolysis gave *N*-labeled glutamic acid. **(a)** On the basis of this information alone, what structures are possible for glutathione? **(b)** Synthetic experiments have shown that the second dipeptide has the following structure:



What is the structure of glutathione?

PRACTICE PROBLEM 24.9

Give the amino acid sequence of the following polypeptides using only the data given by partial acidic hydrolysis:

(a) S, O, P, T $\xrightarrow[H_2O]{H_3O^+}$ ST + TO + PS **(b)** A, R, C, V, L $\xrightarrow[H_2O]{H_3O^+}$ AC + CR + RV + LA

24.6 EXAMPLES OF POLYPEPTIDE AND PROTEIN PRIMARY STRUCTURE

- The covalent structure of a protein or polypeptide is called its **primary structure** (Fig. 24.5).

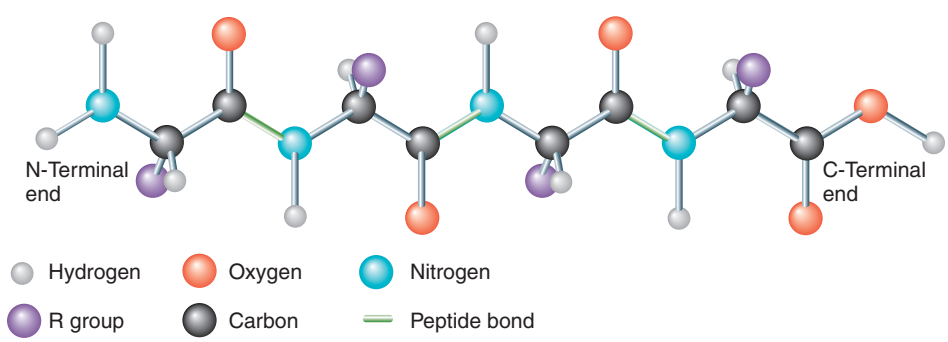


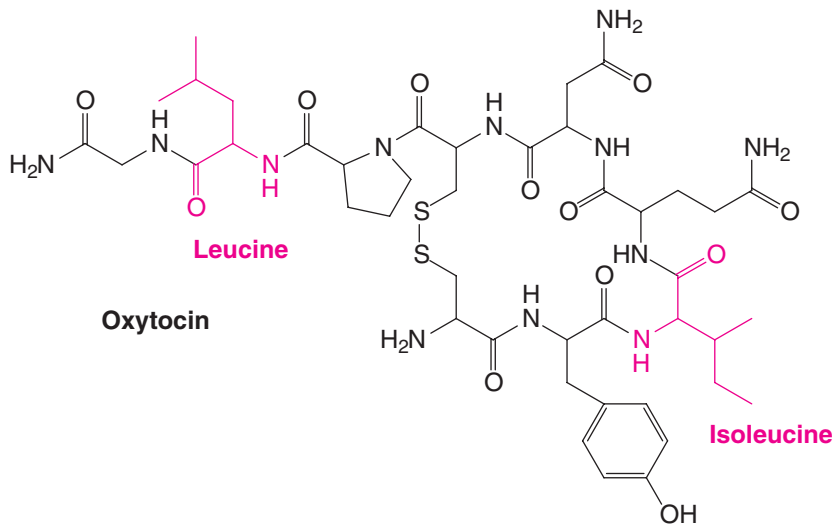
FIGURE 24.5 A representation of the primary structure of a tetrapeptide.

Using the techniques we described, chemists have had remarkable success in determining the primary structures of polypeptides and proteins. The compounds described in the following pages are important examples.

Vincent du Vigneaud of Cornell Medical College synthesized oxytocin and vasopressin in 1953; he received the Nobel Prize in Chemistry in 1955.

24.6A Oxytocin and Vasopressin

Oxytocin and vasopressin (Fig. 24.6) are two rather small polypeptides with strikingly similar structures (where oxytocin has leucine, vasopressin has arginine, and where oxytocin has isoleucine, vasopressin has phenylalanine). In spite of the similarity of their



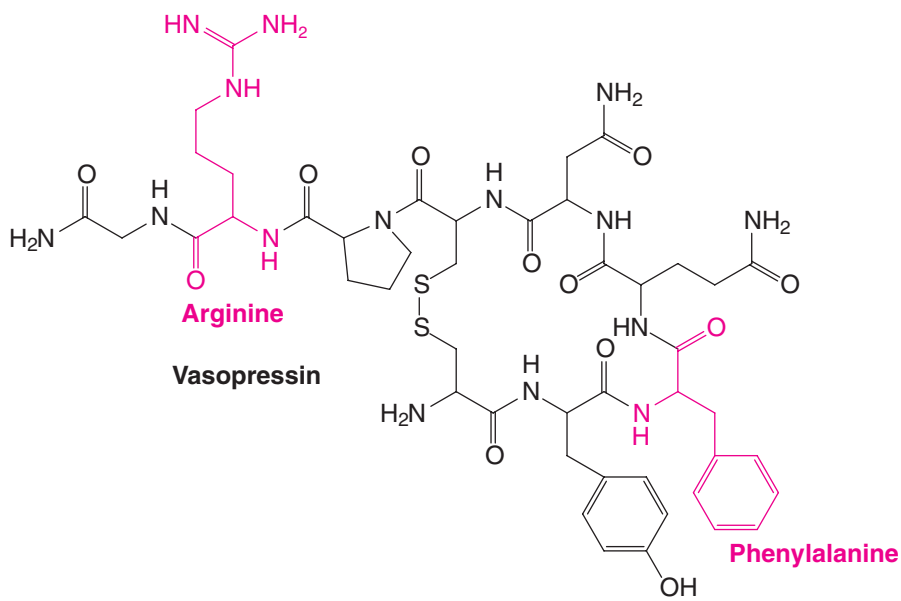


FIGURE 24.6 The structures of oxytocin and vasopressin. Amino acid residues that differ between them are shown in red.

amino acid sequences, these two polypeptides have quite different physiological effects. Oxytocin occurs only in the female of a species and stimulates uterine contractions during childbirth. Vasopressin occurs in males and females; it causes contraction of peripheral blood vessels and an increase in blood pressure. Its major function, however, is as an *antidiuretic;* physiologists often refer to vasopressin as an *antidiuretic hormone.*

The structures of oxytocin and vasopressin also illustrate the importance of the disulfide linkage between cysteine residues (Section 24.2A) in the overall primary structure of a polypeptide. In these two molecules this disulfide linkage leads to a cyclic structure.

PRACTICE PROBLEM 24.10 Treating oxytocin with certain reducing agents (e.g., sodium in liquid ammonia) brings about a single chemical change that can be reversed by air oxidation. What chemical changes are involved?

24.6B Insulin

Insulin, a hormone secreted by the pancreas, regulates glucose metabolism. Insulin deficiency in humans is the major problem in diabetes mellitus.

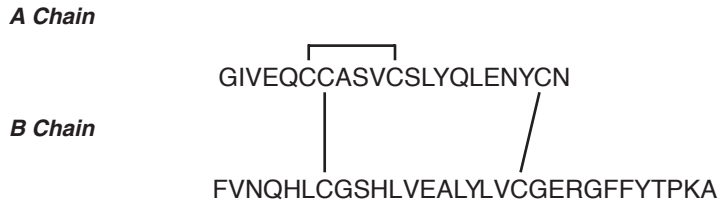


FIGURE 24.7 The amino acid sequence of bovine insulin. Lines between chains indicate disulfide linkages.

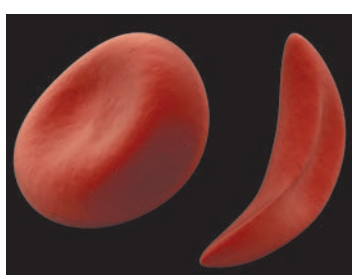
The amino acid sequence of bovine insulin (Fig. 24.7) was determined by Sanger in 1953 after 10 years of work. Bovine insulin has a total of 51 amino acid residues in two polypeptide chains, called the A and B chains. These chains are joined by two disulfide linkages. The A chain contains an additional disulfide linkage between cysteine residues at positions 6 and 11.

Human insulin differs from bovine insulin at only three amino acid residues: Threonine replaces alanine once in the A chain (residue 8) and once in the B chain (residue 30), and isoleucine replaces valine once in the A chain (residue 10). Insulins from most mammals have similar structures.

THE CHEMISTRY OF... Sickle-Cell Anemia

The genetically based disease sickle-cell anemia results from a single amino acid error in the β chain of hemoglobin. In normal hemoglobin, position 6 has a glutamic acid residue, whereas in sickle-cell hemoglobin position 6 is occupied by valine.

Red blood cells (erythrocytes) containing hemoglobin with this amino acid residue error tend to become crescent shaped ("sickle") when the partial pressure of oxygen is low, as it is in venous blood. These distorted cells are more difficult for the heart to pump through small capillaries. They may even block capillaries by clumping together; at other times the red cells may even split open. Children who inherit this genetic trait from both parents suffer from a severe form of the disease and usually do not live past the age of two. Children who inherit the disease from only one parent generally have a much milder form. Sickle-cell anemia arose among the populations of central and western Africa where, ironically, it may have had a beneficial effect. People with a mild form of the disease are far less susceptible to malaria than those with normal hemoglobin. Malaria, a disease caused by an infectious microorganism, is especially prevalent in central and western Africa. Mutational changes such as those that give rise to sickle-cell anemia are very common. Approximately 150 different types of mutant hemoglobin have been detected in humans; fortunately, most are harmless.



Ingram Publishing/SuperStock

Normal (left) and sickled (right) red blood cells viewed with a scanning electron microscope at 18,000× magnification.

24.6C Other Polypeptides and Proteins

Successful sequential analyses have now been achieved with thousands of other polypeptides and proteins, including the following:

1. **Bovine ribonuclease.** This enzyme, which catalyzes the hydrolysis of ribonucleic acid (Chapter 25), has a single chain of 124 amino acid residues and four intrachain disulfide linkages.
2. **Human hemoglobin.** There are four peptide chains in this important oxygen-carrying protein. Two identical α chains have 141 residues each, and two identical β chains have 146 residues each.
3. **Bovine trypsinogen and chymotrypsinogen.** These two digestive enzyme precursors have single chains of 229 and 245 residues, respectively.
4. **Gamma globulin.** This immunoprotein has a total of 1320 amino acid residues in four chains. Two chains have 214 residues each; the other two have 446 each.
5. **p53, an anticancer protein.** The protein called p53 (the p stands for protein), consisting of 393 amino acid residues, has a variety of cellular functions, but the most

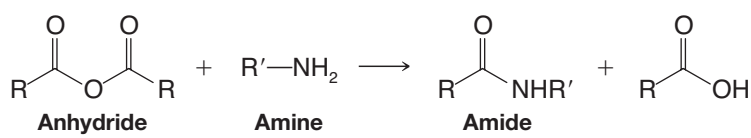
important ones involve controlling the steps that lead to cell growth. It acts as a **tumor suppressor** by halting abnormal growth in normal cells, and by doing so it prevents cancer. Discovered in 1979, p53 was originally thought to be a protein synthesized by an oncogene (a gene that causes cancer). Research has shown, however, that the form of p53 originally thought to have this cancer-causing property was a mutant form of the normal protein. The unmutated (or *wild type*) p53 apparently coordinates a complex set of responses to changes in DNA that could otherwise lead to cancer. When p53 becomes mutated, it no longer provides the cell with its cancer-preventing role; it apparently does the opposite, by acting to increase abnormal growth.

More than half of the people diagnosed with cancer each year have a mutant form of p53 in their cancers. Different forms of cancer have been shown to result from different mutations in the protein, and the list of cancer types associated with mutant p53 includes cancers of most of the body parts: brain, breast, bladder, cervix, colon, liver, lung, ovary, pancreas, prostate, skin, stomach, and so on.

6. ***Ras* proteins.** *Ras* proteins are modified proteins associated with cell growth and the cell's response to insulin. They belong to a class of proteins called prenylated proteins, in which lipid groups derived from isoprenoid biosynthesis (Special Topic E, *WileyPLUS*) are appended as thioethers to C-terminal cysteine residues. Certain mutated forms of *ras* proteins cause oncogenic changes in various eukaryotic cell types. One effect of prenylation and other lipid modifications of proteins is to anchor these proteins to cellular membranes. Prenylation may also assist with molecular recognition of prenylated proteins by other proteins.*

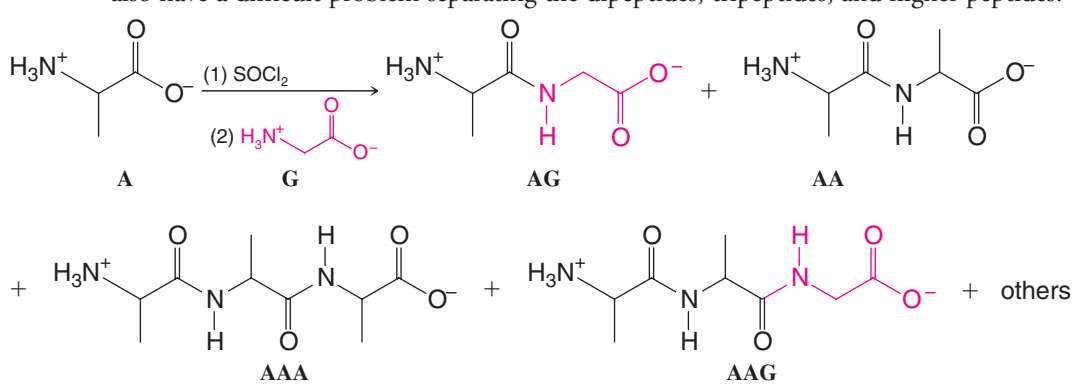
24.7 POLYPEPTIDE AND PROTEIN SYNTHESIS

We saw in Chapter 17 that the synthesis of an amide linkage is a relatively simple one. We must first "activate" the carboxyl group of an acid by converting it to an anhydride or acid chloride and then allow it to react with an amine:



Anhydride **Amine** **Amide**

The problem becomes somewhat more complicated, however, when both the acid group and the amino group are present in the same molecule, as they are in an amino acid, and especially when our goal is the synthesis of a naturally occurring polyamide where the sequence of different amino acids is all important. Let us consider, as an example, the synthesis of the simple dipeptide alanylglycine, AG. We might first activate the carboxyl group of alanine by converting it to an acid chloride, and then we might allow it to react with glycine. Unfortunately, however, we cannot prevent alanyl chloride from reacting with itself. So our reaction would yield not only AG but also AA. It could also lead to AAA and AAG, and so on. The yield of our desired product would be low, and we would also have a difficult problem separating the dipeptides, tripeptides, and higher peptides.



A **G** **AG** **AA**

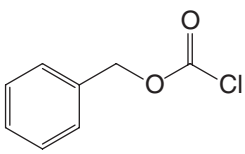
AAA **AAG**

*See Gelb, M. H., "Modification of Proteins by Prenyl Groups," in *Principles of Medical Biology*, Vol. 4 (Bittar, E. E., and Bittar, N., eds.), JAI Press: Greenwich, CT, 1995; Chapter 14, pp. 323–333.

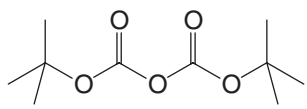
24.7A Protecting Groups

The solution to this problem is to "protect" the amino group of the first amino acid before we activate it and allow it to react with the second. By protecting the amino group, we mean that we must convert it to some other group of low nucleophilicity—*one that will not react with a reactive acyl derivative.* The **protecting group** must be carefully chosen because after we have synthesized the amide linkage between the first amino acid and the second, we will want to be able to remove the protecting group without disturbing the new amide bond.

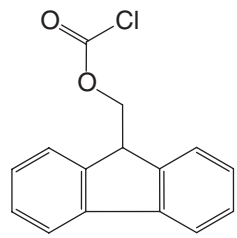
A number of reagents have been developed to meet these requirements. Three that are often used are benzyl chloroformate, di-*tert*-butyl dicarbonate (sometimes abbreviated Boc_2O, where Boc stands for *tert*-butyloxycarbonyl), and 9-fluorenylmethyl chloroformate:



Benzyl chloroformate



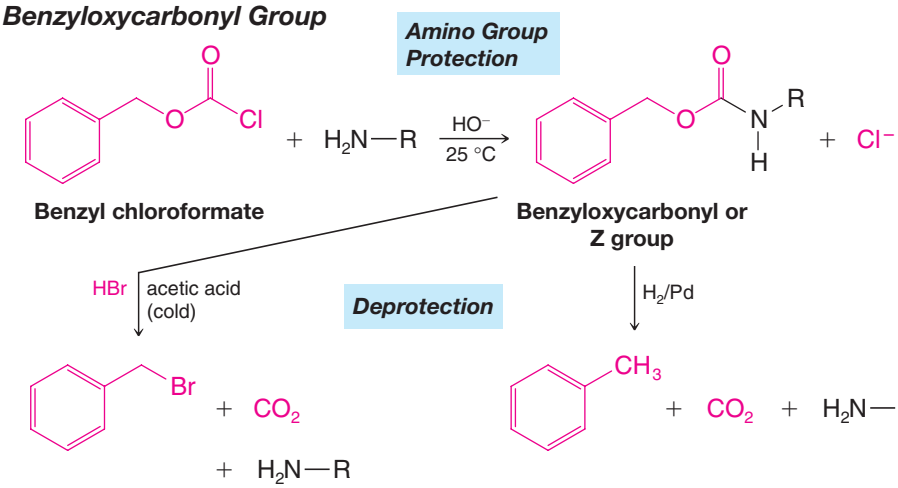
Di-*tert*-butyl dicarbonate (Boc_2O)



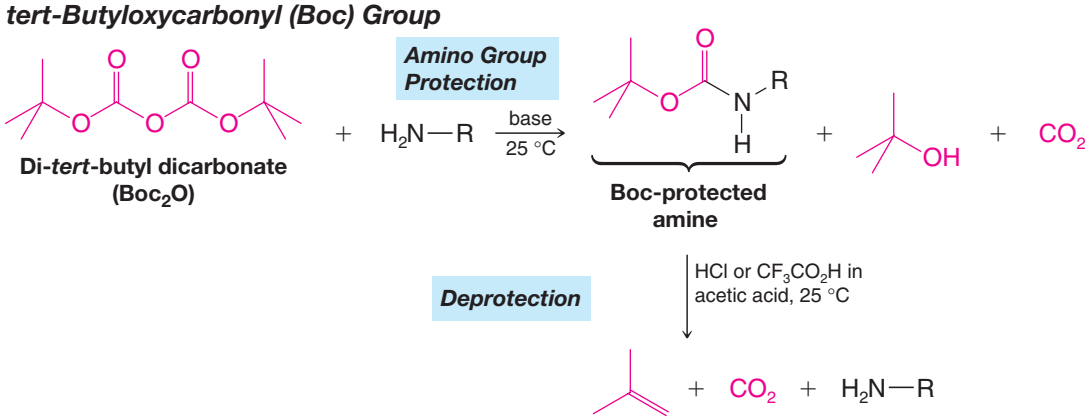
9-Fluorenylmethyl chloroformate

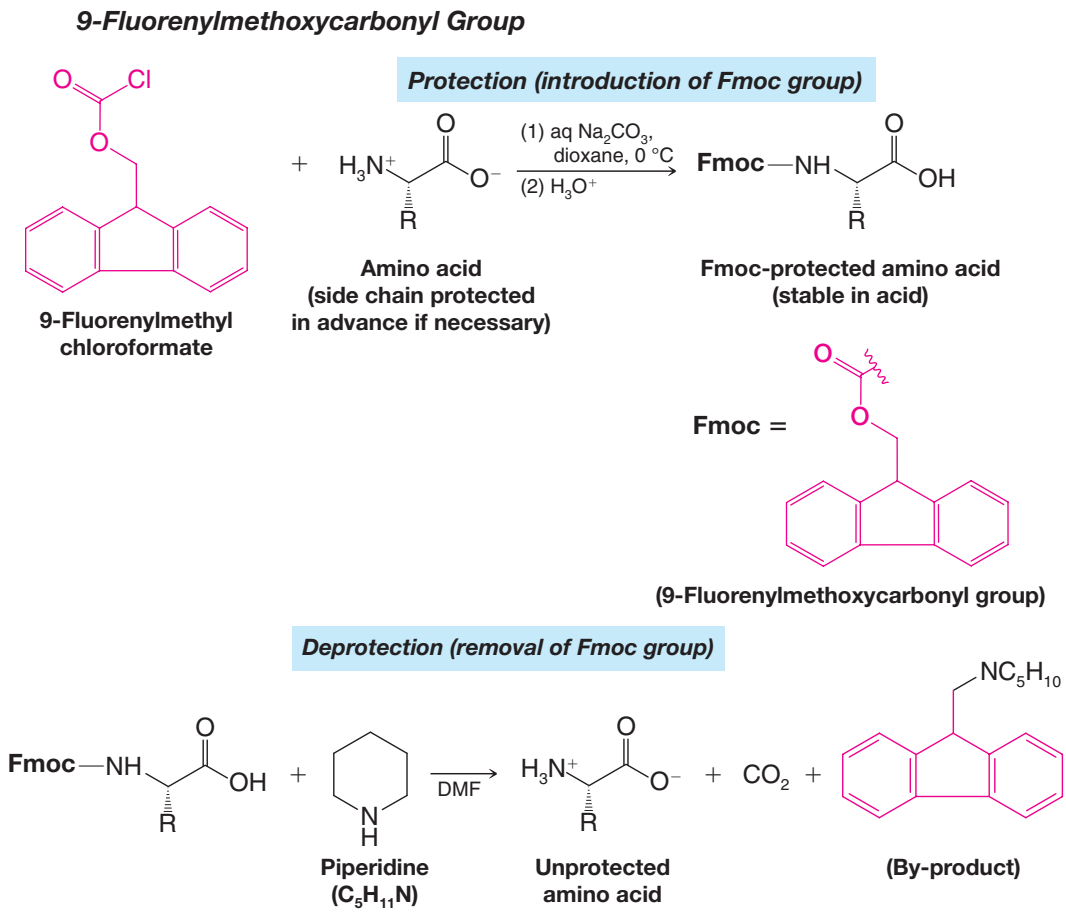
All three reagents react with the amine to block it from further acylation. These derivations, however, are types that allow removal of the protecting group under conditions that do not affect peptide bonds. The benzyloxycarbonyl group (abbreviated Z) can be removed with catalytic hydrogenation or cold HBr in acetic acid. The *tert*-butyloxycarbonyl group can be removed with trifluoroacetic acid (CF_3CO_2H) in acetic acid. The 9-fluorenylmethoxycarbonyl (Fmoc) group is stable under acid conditions but can be removed under mild basic conditions using piperidine (a secondary amine).

Benzyloxycarbonyl Group

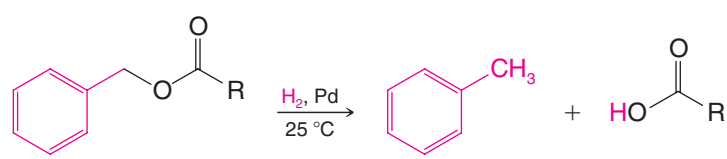


***tert*-Butyloxycarbonyl (Boc) Group**





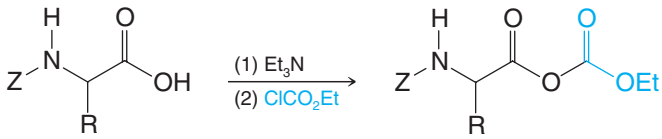
The easy removal of the Z and Boc groups in acidic media results from the exceptional stability of the carbocations that are formed initially. The benzyloxycarbonyl group gives a benzyl carbocation; the *tert*-butyloxycarbonyl group yields, initially, a *tert*-butyl cation. Removal of the benzyloxycarbonyl group with hydrogen and a catalyst depends on the fact that benzyl–oxygen bonds are weak and subject to hydrogenolysis at low temperatures, resulting in methylbenzene (toluene) as one product:



PRACTICE PROBLEM 24.11 What classes of reactions are involved in the cleavage of the Fmoc group with piperidine, leading to the unprotected amino acid and the fluorene by-product? Write mechanisms for these reactions.

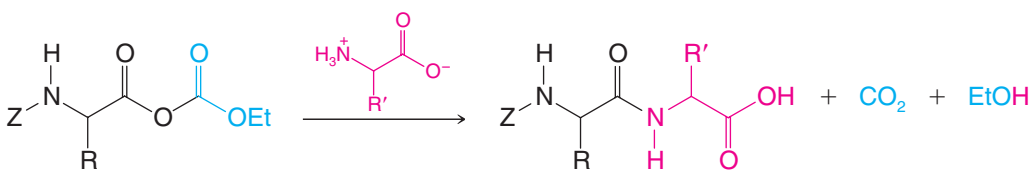
24.7B Activation of the Carboxyl Group

Perhaps the most obvious way to activate a carboxyl group is to convert it to an acyl chloride. This method was used in early peptide syntheses, but acyl chlorides are actually more reactive than necessary. As a result, their use leads to complicating side reactions. A much better method is to convert the carboxyl group of the "protected" amino acid to a mixed anhydride using ethyl chloroformate,  , as shown on the next page:



"Mixed anhydride"

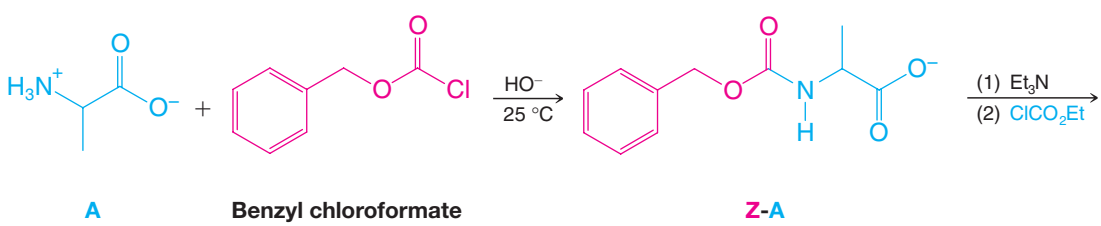
The mixed anhydride can then be used to acylate another amino acid and form a peptide linkage:



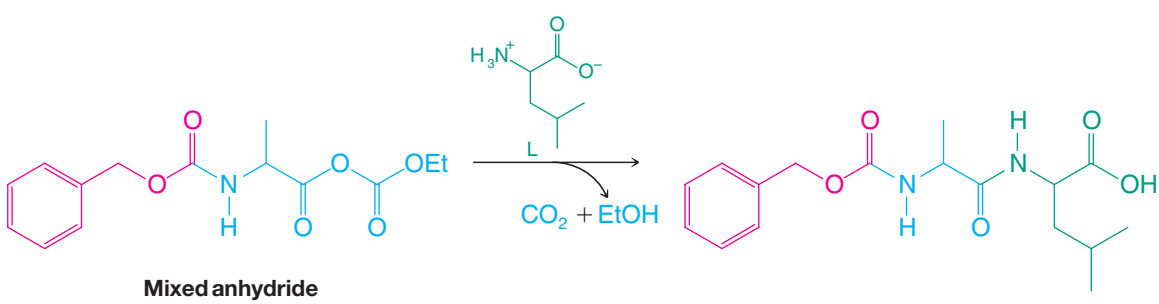
Diisopropylcarbodiimide and dicyclohexylcarbodiimide (Section 17.8E) can also be used to activate the carboxyl group of an amino acid. In Section 24.7D we shall see how diisopropylcarbodiimide is used in an automated peptide synthesis.

24.7C Peptide Synthesis

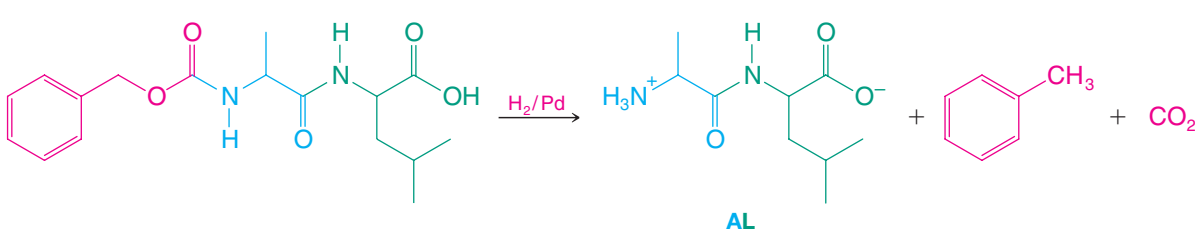
Let us examine now how we might use these reagents in the preparation of the simple dipeptide AL. The principles involved here can, of course, be extended to the synthesis of much longer polypeptide chains.



A **Benzyl chloroformate** **Z-A**



Mixed anhydride of Z-A **Z-AL**



AL

PRACTICE PROBLEM 24.12

Show all steps in the synthesis of GVA using the *tert*-butyloxycarbonyl (Boc) group as a protecting group.

PRACTICE PROBLEM 24.13 The synthesis of a polypeptide containing lysine requires the protection of both amino groups. **(a)** Show how you might do this in a synthesis of the dipeptide KI using the benzyloxycarbonyl group as a protecting group. **(b)** The benzyloxycarbonyl group can also be used to protect the guanidino group, —NHC(=NH)—NH_2, of arginine. Show a synthesis of the dipeptide RA.

PRACTICE PROBLEM 24.14 The terminal carboxyl groups of glutamic acid and aspartic acid are often protected through their conversion to benzyl esters. What mild method could be used for removal of this protecting group?

24.7D Automated Peptide Synthesis

R. B. MERRIFIELD, received the 1984 Nobel Prize in Chemistry for development of a method for solid-phase peptide synthesis.

The methods that we have described thus far have been used to synthesize a number of polypeptides, including ones as large as insulin. They are extremely time-consuming and tedious, however. One must isolate the peptide and purify it by lengthy means at almost every stage. Furthermore, significant loss of the peptide can occur with each isolation and purification stage. The development of a procedure by R. B. Merrifield (Rockefeller University, dec. 2005) for automating this process was therefore a breakthrough in peptide synthesis. Merrifield's method, for which he received the 1984 Nobel Prize in Chemistry, is called **solid-phase peptide synthesis (SPPS),** and it hinges on synthesis of the peptide residue by residue while one end of the peptide remains attached to an insoluble plastic bead. Protecting groups and other reagents are still necessary, but because the peptide being synthesized is anchored to a solid support, by-products, excess reagents, and solvents can simply be rinsed away between each synthetic step without need for intermediate purification. After the very last step the polypeptide is cleaved from the polymer support and subjected to a final purification by HPLC. The method works so well that it was developed into an automated process.

Solid-phase peptide synthesis (Fig. 24.8) begins with attachment of the first amino acid by its carboxyl group to the polymer bead, usually with a linker or spacer molecule in between. Each new amino acid is then added by formation of an amide bond between the N-terminal amino group of the peptide growing on the solid support and the new amino acid's carboxyl group. Diisopropylcarbodiimide (similar in reactivity to DCC, Section 17.8E) is used as the amide bond-forming reagent. To prevent undesired reactions as each new residue is coupled, a protecting group is used to block the amino group of the residue being added. Once the new amino acid has been coupled to the growing peptide and before the next residue is added, the protecting group on the new N-terminus is removed, making the peptide ready to begin the next cycle of amide bond formation.

Although Merrifield's initial method for solid-phase peptide synthesis used the Boc group to protect the α-amino group of residues being coupled to the growing peptide, several advantages of the Fmoc group have since made it the group of choice. The reasons have mainly to do with excellent selectivity for removing the Fmoc group in the presence of other protecting groups used to block reactive side chains along the growing peptide and the ability to monitor the progress of the solid-phase synthesis by spectrophotometry as the Fmoc group is released in each cycle.

Let us discuss the choice of protecting groups further. As noted (Section 24.7A), *basic conditions* (piperidine in DMF) are used to remove the Fmoc group. On the other hand, protecting groups for the side chains of the peptide residues are generally blocked with *acid-labile* moieties. The base-labile Fmoc groups and acid-labile side-chain protecting groups are said to be **orthogonal protecting groups** because one set of protecting groups is stable under conditions for removal of the other, and vice versa. Another advantage of Fmoc as compared to Boc groups for protecting the α-amino group of each new residue

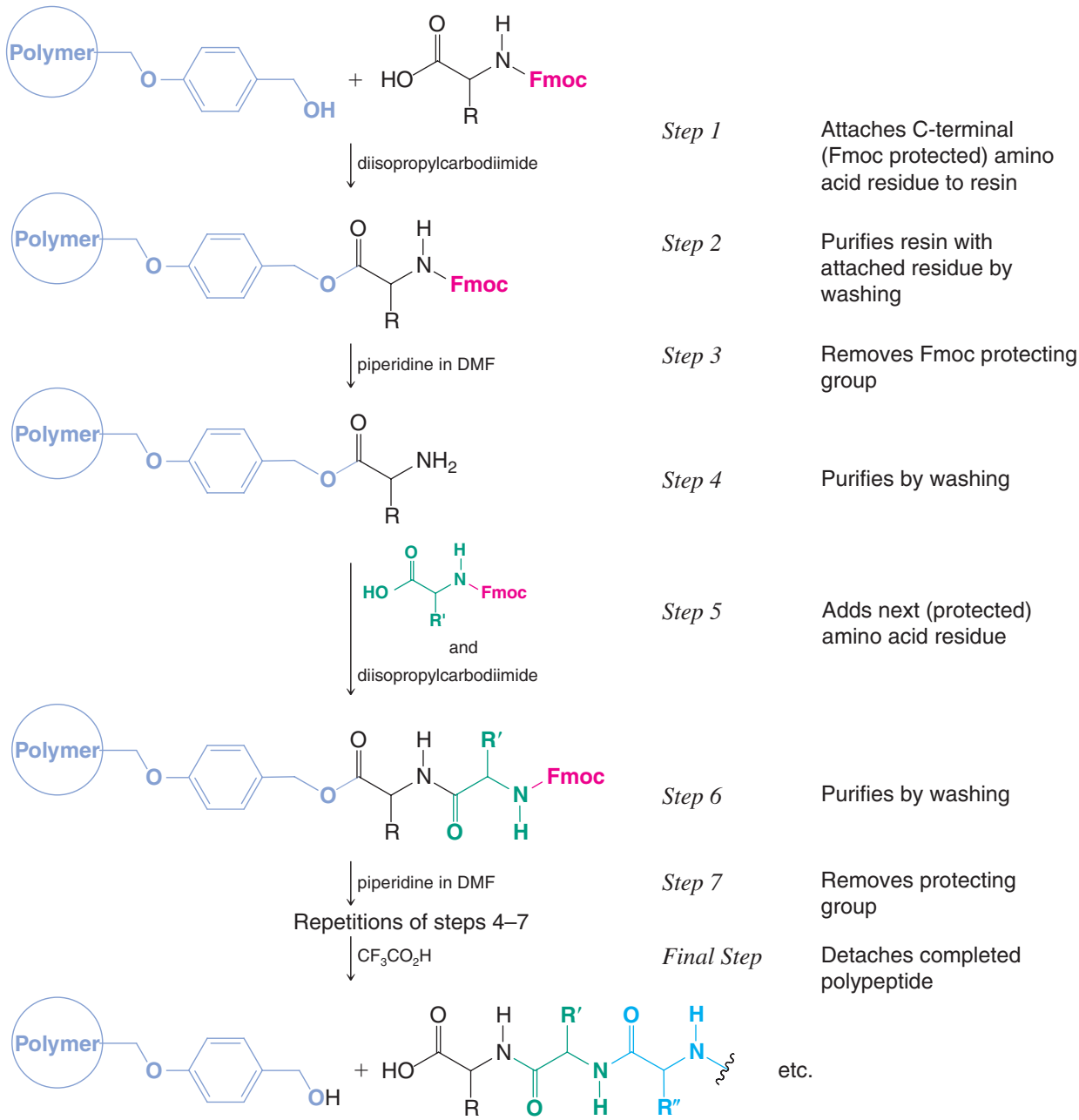


FIGURE 24.8 A method for automated solid-phase peptide synthesis.

is that repetitive application of the acidic conditions to remove Boc groups from each new residue slowly sabotages the synthesis by prematurely cleaving some peptide molecules from the solid support and deprotecting some of the side chains. The basic conditions for Fmoc removal avoid these problematic side reactions.

- The great advantage of solid-phase peptide synthesis is that purification of the peptide at each stage involves simply rinsing the beads of the solid support to wash away excess reagent, by-products, and solvents.
- Furthermore, having the peptide attached to a tangible solid during the synthesis allows all of the steps in the synthesis to be carried out by a machine in repeated cycles.

Automated peptide synthesizers are available that can complete one cycle in 40 min and carry out 45 cycles of unattended operation. Though not as efficient as protein synthesis in the body, where enzymes directed by DNA can catalyze assembly of a protein with 150 amino acids in about 1 min, automated peptide synthesis is a far cry from the

tedious process of manually synthesizing a peptide step after step. A hallmark example of automated peptide synthesis was the synthesis of ribonuclease, a protein with 124 amino acid residues. The synthesis involved 369 chemical reactions and 11,930 automated steps—all carried out without isolating an intermediate. The synthetic ribonuclease not only had the same physical characteristics as the natural enzyme, it possessed the identical biological activity as well. The overall yield was 17%, which means that the average yield of each individual step was greater than 99%.

PRACTICE PROBLEM 24.15 One type of insoluble support used for SPPS is polymer-bound 4-benzyloxybenzyl alcohol, also known as "Wang resin," shown in Fig. 24.8. The 4-benzyloxybenzyl alcohol moiety serves as a linker between the resin backbone and the peptide. After purification, the completed polypeptide can be detached from the resin using trifluoroacetic acid under conditions that are mild enough not to affect the amide linkages. What structural features of the linker make this possible?

PRACTICE PROBLEM 24.16 Outline the steps in the synthesis of the tripeptide KFA using the SPPS procedure.

24.8 SECONDARY, TERTIARY, AND QUATERNARY STRUCTURES OF PROTEINS

We have seen how amide and disulfide linkages constitute the covalent or *primary structure* of proteins. Of equal importance in understanding how proteins function is knowledge of the way in which the peptide chains are arranged in three dimensions. The secondary and tertiary structures of proteins are involved here.

24.8A Secondary Structure

- The **secondary structure** of a protein is defined by the local conformation of its polypeptide backbone.
- Secondary structures are specified in terms of regular folding patterns called **α helices, β sheets**, and **coil** or **loop conformations**.

To understand how these interactions occur, let us look first at what X-ray crystallographic analysis has revealed about the geometry at the peptide bond itself.

- Peptide bonds tend to assume a geometry such that six atoms of the amide linkage are coplanar (Fig. 24.9).

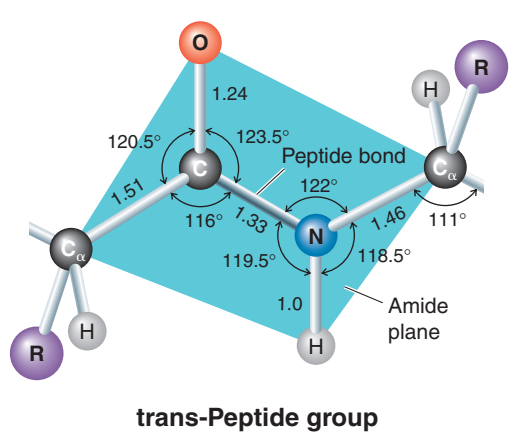
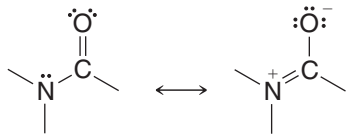
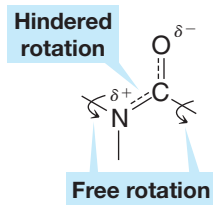


FIGURE 24.9 The geometry and bond lengths (in angstroms, Å) of the peptide linkage. The six enclosed atoms tend to be coplanar and assume a "transoid" arrangement. (Reprinted with permission of John Wiley & Sons, Inc., from Voet, D. and Voet, J. G., *Biochemistry*, Second Edition. © 1995 Voet, D. and Voet, J. G.)

The carbon–nitrogen bond of the amide linkage is unusually short, indicating that resonance contributions of the type shown here are important:



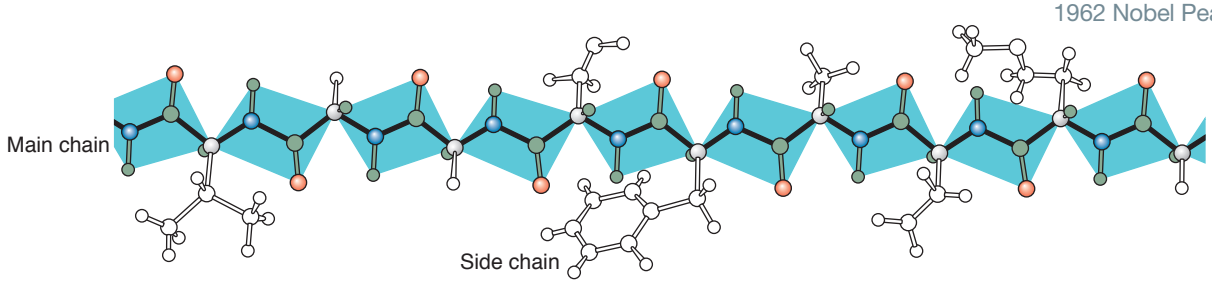
Contributing resonance structures for an amide



Hybrid resonance structure for an amide

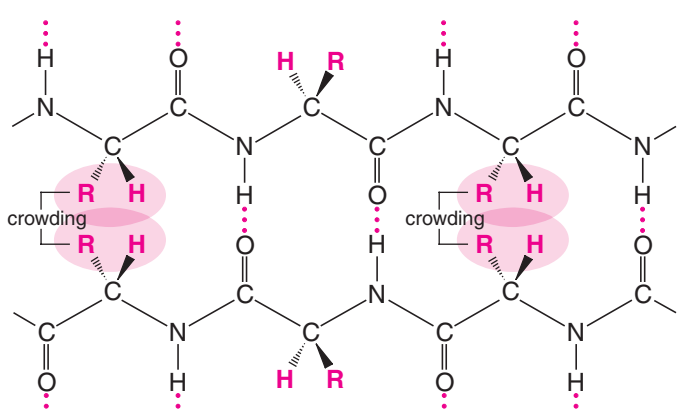
- The amide carbon–nitrogen bond, consequently, has considerable double-bond character (~40%), and rotations of groups about this bond are severely hindered.
- Rotations of groups attached to the amide nitrogen and the carbonyl carbon are relatively free, however, and these rotations allow peptide chains to form different conformations.

A transoid arrangement of groups around the relatively rigid amide bond would cause the side-chain R groups to alternate from side to side of a single fully extended peptide chain:



Calculations show that such a polypeptide chain would have a repeat distance (i.e., distance between alternating units) of 7.2 Å.

Fully extended polypeptide chains could hypothetically form a flat-sheet structure, with each alternating amino acid in each chain forming two hydrogen bonds with an amino acid in the adjacent chain:



Hypothetical flat-sheet structure (not formed because of steric hindrance)

However, this structure does not exist in naturally occurring proteins because of the crowding that would exist between R groups. If such a structure did exist, it would have the same repeat distance as the fully extended peptide chain, that is, 7.2 Å.

Two American scientists, Linus Pauling and Robert B. Corey, were pioneers in the X-ray analysis of proteins. Beginning in 1939, Pauling and Corey initiated a long series of studies of the conformations of peptide chains. At first, they used crystals of single amino acids, then dipeptides and tripeptides, and so on. Moving on to larger and larger molecules and using the precisely constructed molecular models, they were able to understand the secondary structures of proteins for the first time. Pauling won the 1954 Nobel Prize in Chemistry and the 1962 Nobel Peace Prize.

- Many proteins incorporate a ***β* sheet** or ***β* configuration** (Fig. 24.10).

In a β sheet structure, slight bond rotations from one planar amide group to the next relieve the steric strain from small- and medium-sized R groups. This allows amide groups on adjacent polypeptide segments to form hydrogen bonds between the chains (see Fig. 24.10). The β sheet structure has a repeat distance of 7.0 Å between amide groups in a chain. The predominant secondary structure in silk fibroin (48% glycine and 38% serine and alanine residues) is the β sheet.

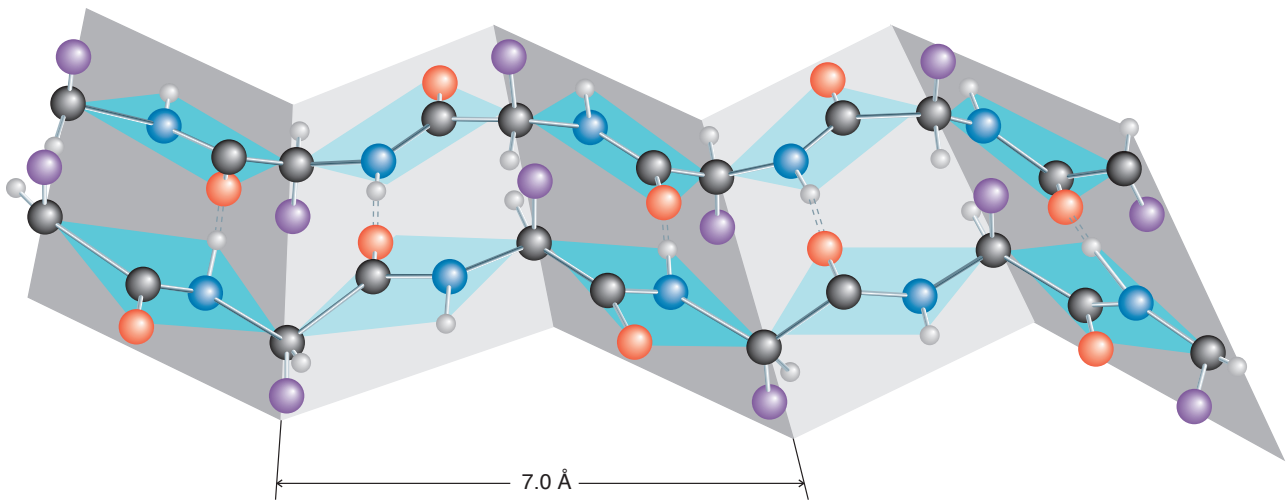


FIGURE 24.10 The β sheet or β configuration of a protein. (Illustration, Irving Geis. Image from the Irving Geis Collection, HHMI. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission.)

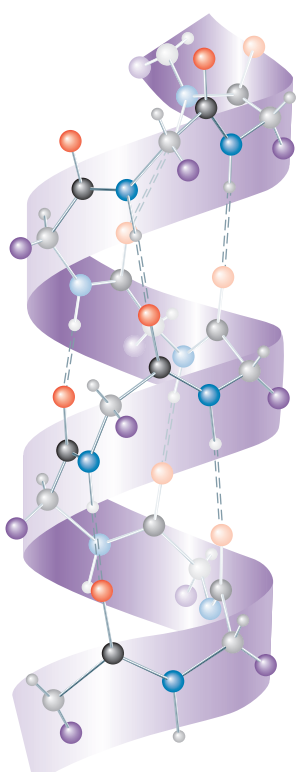


FIGURE 24.11 A representation of the α-helical structure of a polypeptide. Hydrogen bonds are denoted by dashed lines. (Illustration, Irving Geis. Image from the Irving Geis Collection, HHMI. Rights owned by Howard Hughes Medical Institute. Not to be reproduced without permission.)

- The ***α* helix** is also a very important secondary structure in proteins (Fig. 24.11).

The α helix of a polypeptide is right-handed with 3.6 amino acid residues per turn. Each amide group in the chain has a hydrogen bond to an amide group at a distance of three amino acid residues in either direction, and the R groups all extend away from the axis of the helix. The repeat distance of the α helix is 5.4 Å.

The α-helical structure is found in many proteins; it is the predominant structure of the polypeptide chains of fibrous proteins such as *myosin*, the protein of muscle, and of *α-keratin*, the protein of hair, unstretched wool, and nails.

Helices and pleated sheets account for only about one-half of the structure of the average globular protein. The remaining polypeptide segments have what is called a **coil** or **loop conformation**. These nonrepetitive structures are not random; they are just more difficult to describe. Globular proteins also have stretches, called **reverse turns** or ***β* bends**, where the polypeptide chain abruptly changes direction. These often connect successive strands of β sheets and almost always occur at the surface of proteins.

Figure 24.12 shows the structure of the enzyme human carbonic anhydrase, based on X-ray crystallographic data. Segments of α helix (magenta) and β sheets (yellow) intervene between reverse turns and nonrepetitive structures (blue and white, respectively).

- The locations of the side chains of amino acids of globular proteins are usually those that we would expect from their polarities:

1. Residues with **nonpolar, hydrophobic side chains**, such as *valine, leucine, isoleucine, methionine, and phenylalanine,* are almost always found in the interior of the protein, out of contact with the aqueous solvent. (These hydrophobic interactions are largely responsible for the tertiary structure of proteins that we discuss in Section 24.8B.)
2. Side chains of **polar residues with positive or negative charges**, such as *arginine, lysine, aspartic acid, and glutamic acid,* are usually on the surface of the protein in contact with the aqueous solvent.

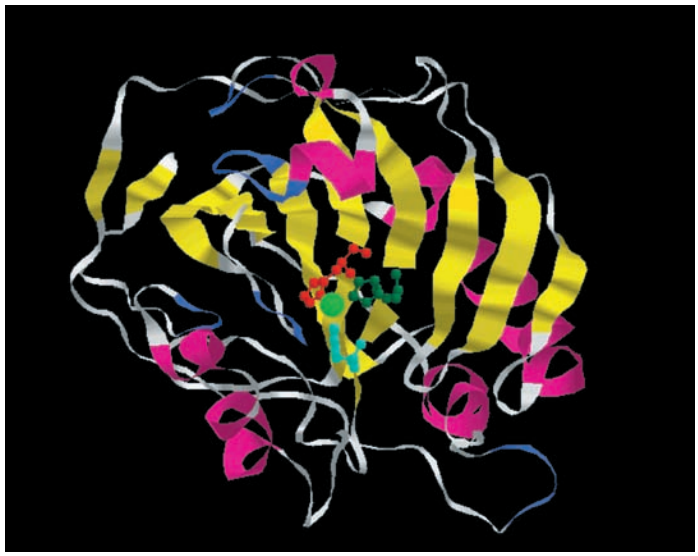


FIGURE 24.12 The structure of the enzyme human carbonic anhydrase, based on X-ray crystallographic data. Alpha helices are shown in magenta and strands of β sheets are yellow. Turns are shown in blue and random coils are white. The side chains of three histidine residues (shown in red, green, and cyan) coordinate with a zinc atom (light green). Not obvious from this image is the interesting fact that the C-terminus is tucked through a loop of the polypeptide chain, making carbonic anhydrase a rare example of a native protein in which the polypeptide chain forms a knot. (PDB ID CA2, *http://www.pdb.org*. Eriksson, Jones, Liljas, *Proteins: Structure, Function and Genetics*, Volume 4, Issue 4, 1988, pp. 274–282.)

3. **Uncharged polar side chains**, such as those of *serine, threonine, asparagine, glutamine, tyrosine, and tryptophan,* are most often found on the surface, but some of these are found in the interior as well. When they are found in the interior, they are virtually all hydrogen bonded to other similar residues. Hydrogen bonding apparently helps neutralize the polarity of these groups.

Certain peptide chains assume what is called a **random coil arrangement**, a structure that is flexible, changing, and statistically random. Synthetic polylysine, for example, exists as a random coil and does not normally form an α helix. At pH 7, the ε-amino groups of the lysine residues are positively charged, and, as a result, repulsive forces between them are so large that they overcome any stabilization that would be gained through hydrogen bond formation of an α helix. At pH 12, however, the ε-amino groups are uncharged and polylysine spontaneously forms an α helix.

The presence of proline or hydroxyproline residues in polypeptide chains produces another striking effect: because the nitrogen atoms of these amino acids are part of five-membered rings, the groups attached by the nitrogen–α carbon bond cannot rotate enough to allow an α-helical structure. Wherever proline or hydroxyproline occur in a peptide chain, their presence causes a kink or bend and interrupts the α helix.

24.8B Tertiary Structure

- The **tertiary structure** of a protein is the overall three-dimensional shape that arises from all of the secondary structures of its polypeptide chain.

Proteins typically have either **globular** or **fibrous** tertiary structures. These tertiary structures do not occur randomly. Under the proper environmental conditions the tertiary structure of a protein occurs in one particular way—a way that is characteristic of that particular protein and one that is often highly important to its function.

Various forces are involved in stabilizing tertiary structures, including the disulfide bonds of the primary structure.

- One characteristic of most proteins is that the folding takes place in such a way as to expose the maximum number of polar (hydrophilic) groups to the aqueous environment and enclose a maximum number of nonpolar (hydrophobic) groups within its interior.

The soluble globular proteins tend to be much more highly folded than fibrous proteins. Myoglobin (Fig. 24.13) is an example of a globular protein. However, fibrous proteins also have a tertiary structure; the α-helical strands of α-keratin, for example, are wound together into a "superhelix." The superhelix makes one complete turn for each 35 turns of the α helix. The tertiary structure does not end here, however. Even the superhelices can be wound together to give a ropelike structure of seven strands.

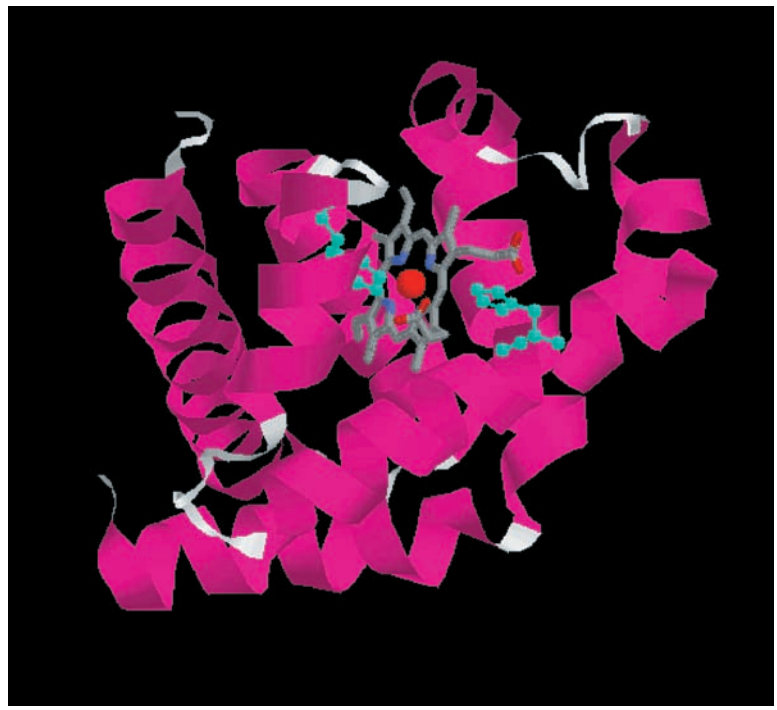
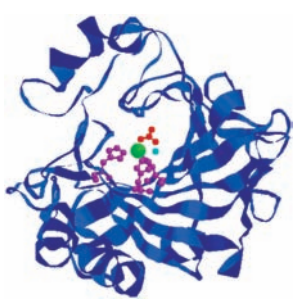


FIGURE 24.13 The three-dimensional structure of myoglobin. The heme ring is shown in gray. The iron atom is shown as a red sphere, and the histidine side chains that coordinate with the iron are shown in cyan. (PDB ID 1MBD, *http://www.pdb.org*. Phillips, S.E., Schoenberg, B.P. Neutron diffraction reveals oxygen-histidine hydrogen bond in oxymyoglobin. *Nature* 292, pp. 81–82, 1981.)

24.8C Quaternary Structure

Many proteins exist as stable and ordered noncovalent aggregates of more than one polypeptide chain. The overall structure of a protein having multiple subunits is called its **quaternary structure**. The quaternary structure of hemoglobin, for example, involves four subunits (see Section 24.12).

24.9 INTRODUCTION TO ENZYMES



Carbonic anhydrase

Carbonic anhydrase is an enzyme that catalyzes the following reaction: $H_2O + CO_2 \rightleftharpoons H_2CO_3$. (PDB ID CA2, *http://www.pdb.org*. Eriksson, Jones, Liljas, *Proteins: Structure, Function and Genetics*, Volume 4, Issue 4, 1988, pp. 274–282.)

- The reactions of cellular metabolism are mediated by remarkable biological catalysts called **enzymes**.

Enzymes have the ability to bring about vast increases in the rates of reactions; in most instances, the rates of enzyme-catalyzed reactions are faster than those of uncatalyzed reactions by factors of 10^6–10^{12}. For living organisms, rate enhancements of this magnitude are important because they permit reactions to take place at reasonable rates, even under the mild conditions that exist in living cells (i.e., approximately neutral pH and a temperature of about 35 °C).

- Enzymes show remarkable **specificity** for their **substrates** and for formation of specific products.

The specificity of enzymes is far greater than that shown by most chemical catalysts. In the enzymatic synthesis of proteins, for example (through reactions that take place on ribosomes, Section 25.5E), polypeptides consisting of well over 1000 amino acid residues are synthesized virtually without error. It was Emil Fischer's discovery, in 1894, of the ability of enzymes to distinguish between α- and β-glycosidic linkages (Section 22.12) that led him to formulate his **lock-and-key hypothesis** for enzyme specificity.

- According to the **lock-and-key hypothesis**, the specificity of an enzyme (the lock) and its substrate (the key) comes from their geometrically complementary shapes.
- In an enzyme-catalyzed reaction, the enzyme and the substrate combine to form an **enzyme–substrate complex**.
- Formation of the enzyme–substrate complex often induces a conformational change in the enzyme called an **induced fit** that allows it to bind the substrate more effectively.

Binding of the substrate can cause certain of its bonds to become strained, and therefore more easily broken. The product of the reaction usually has a different shape from the substrate, and this altered shape, or in some instances the intervention of another molecule, causes the complex to dissociate. The enzyme can then accept another molecule of the substrate, and the whole process is repeated:

$$\text{Enzyme} + \text{substrate} \rightleftharpoons \text{enzyme–substrate complex} \rightleftharpoons \text{enzyme} + \text{product}$$

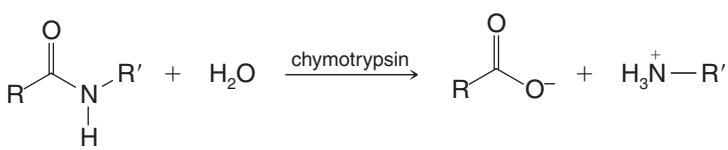
- The place where a substrate binds to an enzyme and where the reaction takes place is called the **active site**.

The noncovalent forces that bind the substrate to the active site are the same forces that account for the conformations of proteins: dispersion forces, electrostatic forces, hydrogen bonding, and hydrophobic interactions. The amino acids located in the active site are arranged so that they can interact specifically with the substrate.

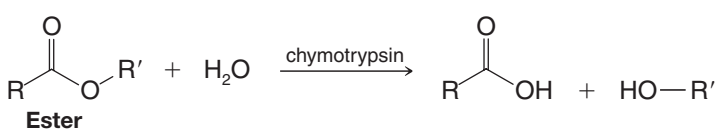
- Reactions catalyzed by enzymes are **stereospecific** because enzymes are chiral.

The specificity of enzymes arises in the way enzymes bind their substrates. An α-glycosidase will only bind the α stereoisomeric form of a glycoside, not the β form. Enzymes that metabolize sugars bind only D sugars; enzymes that synthesize most proteins bind only L amino acids; and so on.

Although enzymes catalyze reactions stereospecifically, they often vary considerably in what is called their **geometric specificity**. By geometric specificity, we mean a specificity that is related to the identities of the chemical groups of the substrates. Some enzymes will accept only one compound as their substrate. Others, however, will accept a range of compounds with similar groups. Carboxypeptidase A, for example, will hydrolyze the C-terminal peptide from all polypeptides as long as the penultimate residue is not arginine, lysine, or proline and as long as the next preceding residue is not proline. Chymotrypsin, a digestive enzyme that catalyzes the hydrolysis of peptide bonds, will also catalyze the hydrolysis of esters. We shall consider its mechanism of hydrolysis in Section 24.11.



Peptide



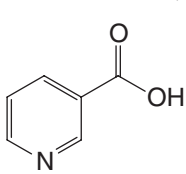
Ester

- A compound that can negatively alter the activity of an enzyme is called an **inhibitor**.

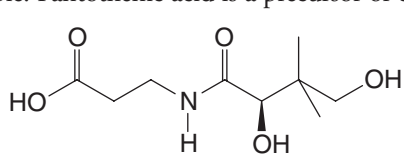
A **competitive inhibitor** is a compound that competes directly with the substrate for the active site. We learned in Section 20.9, for example, that sulfanilamide is a competitive inhibitor for a bacterial enzyme that incorporates *p*-aminobenzoic acid into folic acid.

Some enzymes require the presence of a **cofactor**. The cofactor may be a metal ion as, for example, the zinc atom of human carbonic anhydrase (see the Chemistry of…box, Section 24.10 and Fig. 24.12). Others may require the presence of an organic molecule, such as NAD^+ (Section 14.10), called a **coenzyme**. Coenzymes become chemically changed in the course of the enzymatic reaction. NAD^+ becomes converted to **NADH**. In some enzymes the cofactor is permanently bound to the enzyme, in which case it is called a **prosthetic group**.

Many of the water-soluble vitamins are the precursors of coenzymes. Niacin (nicotinic acid) is a precursor of NAD^+, for example. Pantothenic acid is a precursor of coenzyme A.



Niacin



Pantothenic acid

Certain RNA molecules, called ribozymes, can also act as enzymes. The 1989 Nobel Prize in Chemistry went to Sidney Altman (Yale University) and to Thomas R. Cech (University of Colorado, Boulder) for the discovery of ribozymes.

Helpful Hint

In *WileyPLUS* we have highlighted several coenzymes because they are the "organic chemistry machinery" of some enzymes. For example, see "The Chemistry of ... Pyridoxal Phosphate" and "The Chemistry of ... Thiamine."

24.10 LYSOZYME: MODE OF ACTION OF AN ENZYME

Lysozyme is an enzyme that breaches the cell wall of gram-positive bacteria by hydrolyzing specific acetal linkages in the cell's peptidoglycan polymer, causing lysis and cell death. We shall discuss the mechanism of this reaction below, but first let us consider the structure of lysozyme. The primary structure of lysozyme is shown in Figure 24.14.

Lysozyme's secondary structure includes α-helices at residues 5–15, 24–34, and 88–96; β-sheet involving residues 41–45 and 50–54; and a hairpin turn at residues 46–49. The remaining polypeptide segments of lysozyme have coil or loop formations. Glu-35 and Asp-52 are the amino acid residues directly involved in the hydrolysis reaction catalyzed by lysozyme. A three-dimensional structure of lysozyme is shown in Fig. 24.15. The amino acid residues responsible for its catalytic activity are highlighted in ball-and-stick format (Glu-35 and Asp-52 to the left).

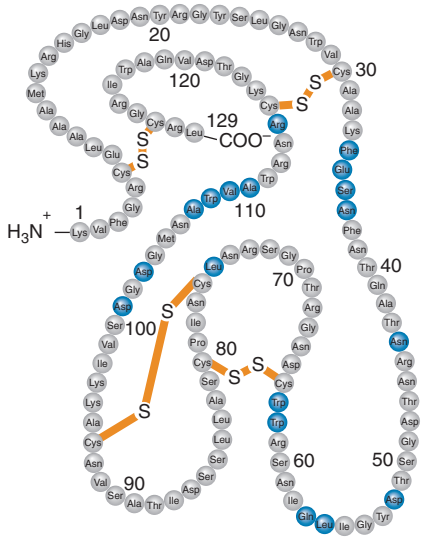


FIGURE 24.14 The primary structure of hen egg white lysozyme. The amino acids that line the substrate-binding pocket are shown in blue. (Reprinted with permission of John Wiley & Sons, Inc. from Voet, D. and Voet, J. G. *Biochemistry*, Second Edition. © 1995 Voet, D. and Voet, J. G.)

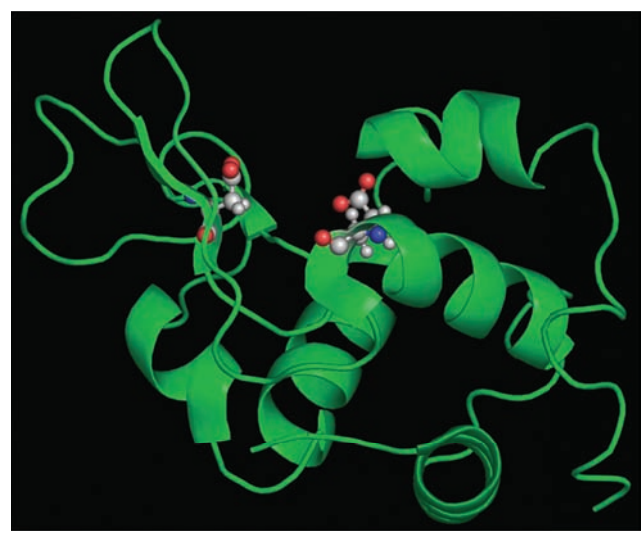


FIGURE 24.15 A ribbon diagram of lysozyme highlighting aspartic acid 52 (left) and glutamic acid 35 (right) in ball-and-stick format. (PDB ID: 1AZF, *http://www.pdb.org*. Lim, K., Nadarajah A., Forsythe, E. L., Pusey, M. L. Locations of bromide ions in tetragonal lysozyme crystals. *Acta Crystallogr.*, Sect. D, 54, pp. 899–904, 1998.)

As mentioned, lysozyme hydrolyzes glycosidic linkages in the peptidoglycan polymer of gram-positive bacterial cell walls. The structure of an oligosaccharide similar to the polysaccharide found in bacterial cell walls is shown in Fig. 24.16. *N*-Acetylglucosamine (NAG) and *N*-acetylmuramic acid (NAM) form alternating repeat units in this polysaccharide.

Lysozyme selectively binds a six-unit segment of the peptidoglycan polymer and hydrolyzes specifically the acetal linkage between rings D and E shown in Fig. 24.16 (NAM and NAG units, respectively).

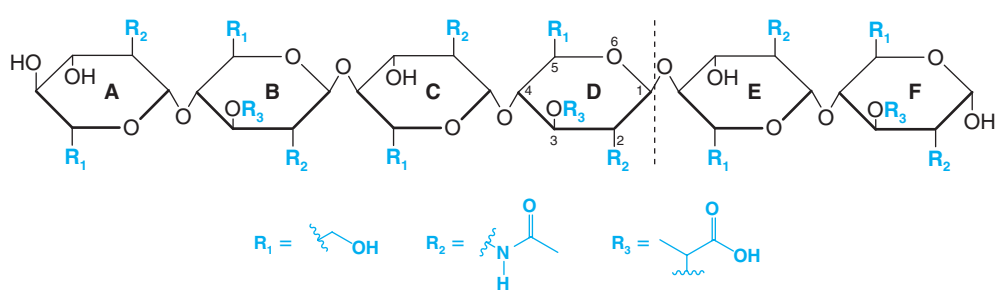
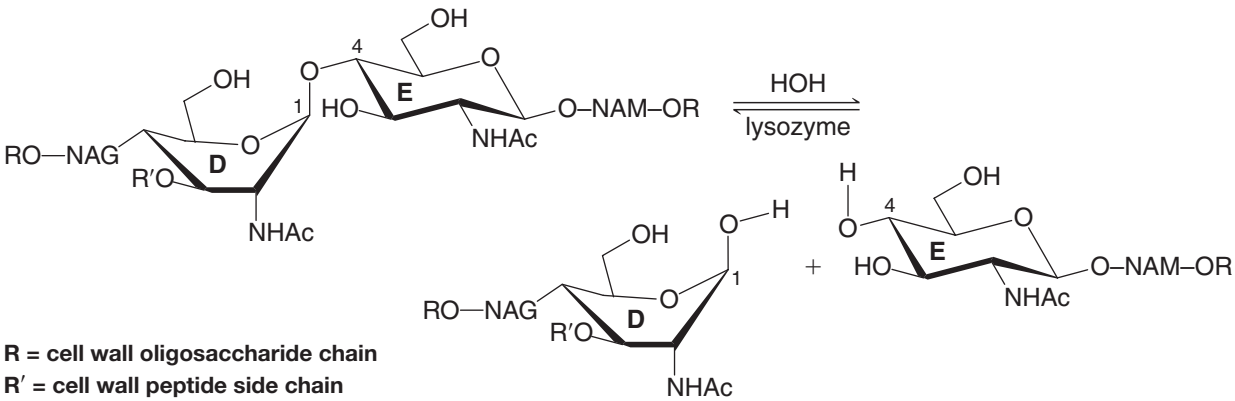


FIGURE 24.16 A hexasaccharide that has the same general structure as the cell wall polysaccharide on which lysozyme acts. Two different amino sugars are present: rings A, C, and E are derived from a monosaccharide called *N*-acetylglucosamine; rings B, D, and F are derived from a monosaccharide called *N*-acetylmuramic acid. When lysozyme acts on this oligosaccharide, hydrolysis takes place and results in cleavage at the glycosidic linkage between rings D and E.

The overall reaction catalyzed by lysozyme is as follows:



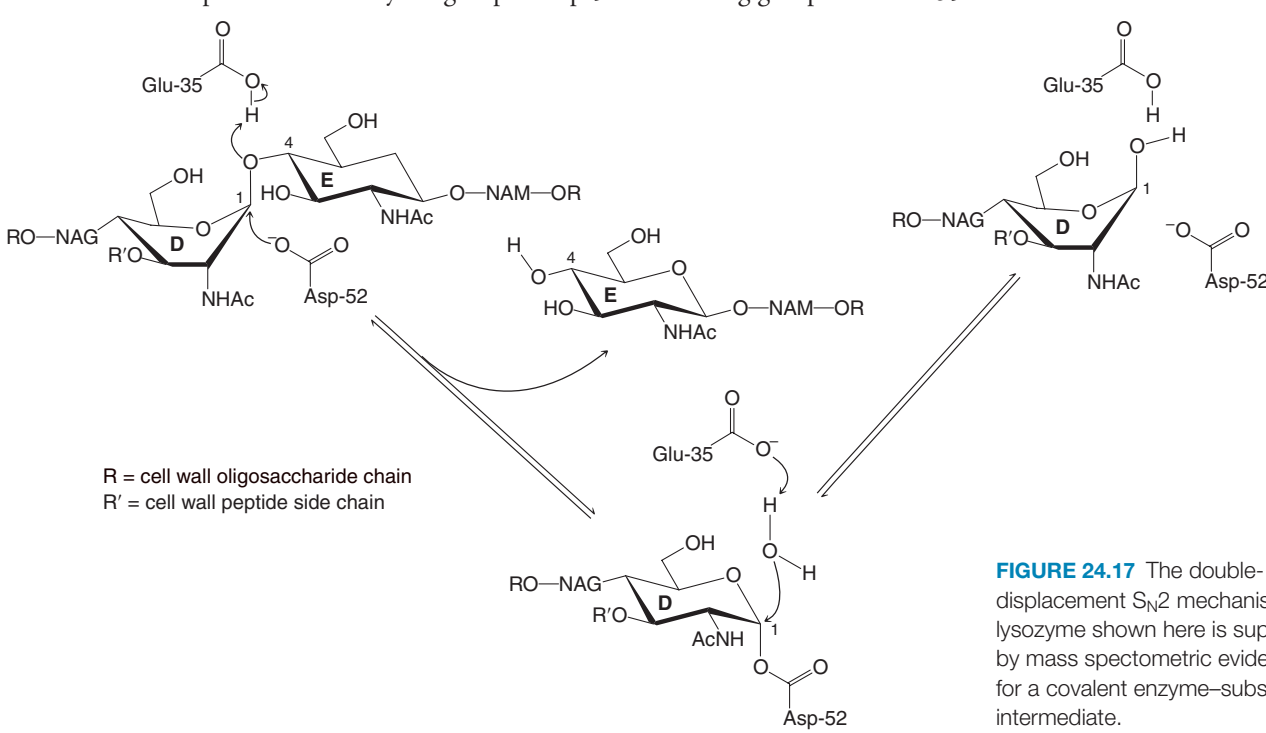
R = cell wall oligosaccharide chain
R′ = cell wall peptide side chain

Lysozyme binds the cell wall substrate in a cleft within its tertiary structure, such that the Glu-35 residue is close to the substrate on one side and Asp-52 is close on the other. Both amino acid residues are positioned in a way that facilitates reaction with the D–E glycosidic linkage of the polysaccharide.

Strong evidence from mass spectrometry suggests that the mechanism of lysozyme involves sequential S_N2 reactions and a covalent enzyme–substrate intermediate (based on work by Stephen Withers and colleagues at the University of British Columbia and elsewhere). Asp-52 acts as the nucleophile in the first step, covalently bonding the substrate to the enzyme. A water molecule acts as a nucleophile in the second step to complete the formation of product and free the substrate from the active site. In both steps, Glu-35 serves as a general acid–base catalyst. The details are as follows.

As lysozyme binds the substrate, the active site cleft closes slightly and C1 of ring D in the oligosaccharide substrate moves downward. The carboxylate group of Asp-52 attacks C1 of ring D from below (Figure 24.17), displacing the ring E C4 oxygen as a leaving group. The ring E C4 oxygen departs as a neutral species because it is protonated concurrently by the carboxylic acid group of Glu-35. The transition state for this S_N2 reaction is presumed to be the point at which ring D is nearly flat during the boat to chair conformational change. This step occurs with inversion, as expected for an S_N2 reaction, and leaves one part of the substrate covalently bound to the enzyme.

In the second step, a water molecule, now in the site formerly occupied by ring E, attacks C1 and displaces the carboxylate group of Asp-52 as a leaving group. The Glu-35



R = cell wall oligosaccharide chain
R′ = cell wall peptide side chain

FIGURE 24.17 The double-displacement S_N2 mechanism for lysozyme shown here is supported by mass spectometric evidence for a covalent enzyme–substrate intermediate.

anion assists as a base by removing a proton from the water molecule as it bonds with C1 of ring D. The entire lysozyme molecule serves as the leaving group. This event also occurs with inversion, liberates the substrate from the active site, and returns lysozyme to readiness for another catalytic cycle. The overall mechanism is shown in Fig. 24.17.

THE CHEMISTRY OF... Carbonic Anhydrase: Shuttling the Protons

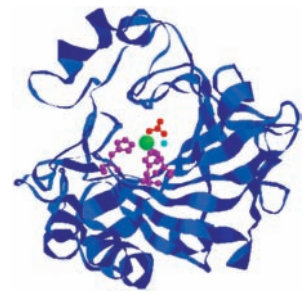
An enzyme called carbonic anhydrase regulates the acidity (pH) of blood and the physiological conditions relating to blood pH. The reaction that carbonic anhydrase catalyzes is the equilibrium conversion of water and carbon dioxide to carbonic acid (H_2CO_3).

$$H_2O + CO_2 \overset{\text{carbonic anhydrase}}{\rightleftharpoons} H_2CO_3 \rightleftharpoons HCO_3^- + H^+$$

The rate at which one breathes, for example, is influenced by one's relative blood acidity. Mountain climbers going to high elevations sometimes take a drug called Diamox (acetazolamide) to prevent altitude sickness. Diamox inhibits carbonic anhydrase, and this, in turn, increases blood acidity. This increased blood acidity stimulates breathing and thereby decreases the likelihood of altitude sickness.

Carbonic anhydrase consists of a chain of 260 amino acids that naturally folds into a specific globular shape. Included in its structure is a cleft or pocket, the active site, where the reactants are converted to products. The protein chain of carbonic anhydrase is shown here as a blue ribbon.

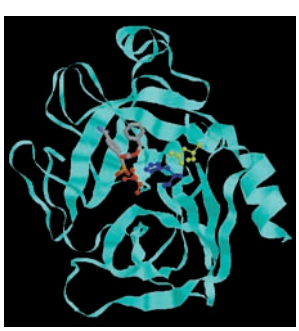
At the active site of carbonic anhydrase a water molecule loses a proton to form a hydroxide (HO^-) ion. This proton is removed by a part of carbonic anhydrase that acts as a base. Ordinarily the proton of a water molecule is not very acidic. However, the Lewis acid–base interaction between a zinc cation at the active site of carbonic anhydrase and the oxygen atom of a water molecule leads to positive charge on the water oxygen atom. This makes the protons of the water molecule more acidic. Removal of one of the protons of the water molecule forms hydroxide, which reacts with a carbon dioxide molecule at the active site to form HCO_3^- (hydrogen carbonate, or bicarbonate). In the structure of carbonic anhydrase shown here (based on X-ray crystallographic data), a bicarbonate ion at the active site is shown in red, the zinc cation at the active site is green, a water molecule is shown in blue, and the basic sites that coordinate with the zinc cation (as Lewis bases) or remove the proton from water to form hydroxide (as Brønsted–Lowry bases) are magenta (these bases are nitrogen atoms from histidine imidazole rings). No hydrogen atoms are shown in any of these species. As you can see, a remarkable orchestration of Lewis and Brønsted–Lowry acid–base reactions is involved in catalysis by carbonic anhydrase.



Carbonic anhydrase

(PDB ID CA2, *http://www.pdb.org.* Eriksson, Jones, Liljas, *Proteins: Structure, Function and Genetics*, Volume 4, Issue 4, pp. 274–282, 1988.)

24.11 SERINE PROTEASES



A serine protease

Chymotrypsin, trypsin, and elastin are digestive enzymes secreted by the pancreas into the small intestine to catalyze the hydrolysis of peptide bonds. These enzymes are all called **serine proteases** because the mechanism for their proteolytic activity (one that they have in common) involves a particular serine residue that is essential for their enzymatic activity. As another example of how enzymes work, we shall examine the mechanism of action of chymotrypsin.

Chymotrypsin is formed from a precursor molecule called chymotrypsinogen, which has 245 amino acid residues. Cleavage of two dipeptide units of chymotrypsinogen produces chymotrypsin. Chymotrypsin folds in a way that brings together histidine at position 57, aspartic acid at position 102, and serine at position 195. Together, these residues constitute what is called the **catalytic triad** of the active site (Fig. 24.18). Near the active

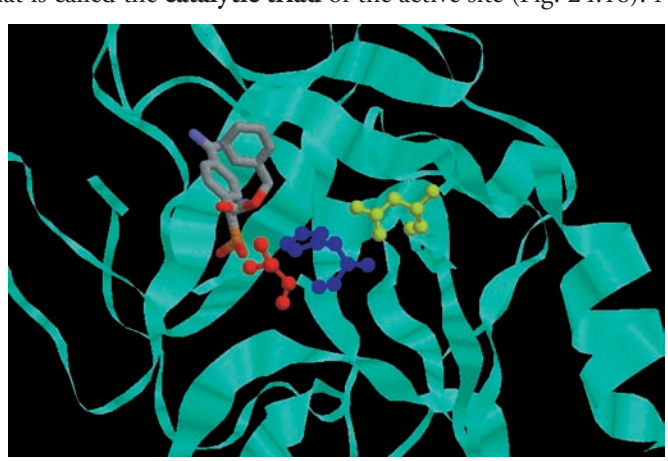


FIGURE 24.18 The catalytic triad in this serine protease (trypsin) is highlighted using the ball-and-stick model format for aspartic acid 52 (yellow-green), histidine 102 (purple), and serine 195 (red). A phosphonate inhibitor bound at the active site is shown in tube format. (This image and that in the margin, PDB ID: 1MAX, http://www.pdb.org. Bertrand, J. A., Oleksyszyn, J., Kam, C. M., Boduszek, B., Presnell, S., Plaskon, R. R., Suddath, F. L., Powers, J. C., Williams, L. D., Inhibition of trypsin and thrombin by amino(4-amidinophenyl)methanephosphonate diphenyl ester derivatives: X-ray structures and molecular models. *Biochemistry* 35, pp. 3147–3155, 1996.)

site is a hydrophobic binding site, a slotlike pocket that preferentially accommodates the nonpolar side chains of Phe, Tyr, and Trp.

After chymotrypsin has bound its protein substrate, the serine residue at position 195 is ideally situated to attack the acyl carbon of the peptide bond (Fig. 24.19). This serine residue is made more nucleophilic by transferring its proton to the imidazole nitrogen of the histidine residue at position 57. The imidazolium ion that is formed is stabilized by the polarizing effect of the carboxylate ion of the aspartic acid residue at position 102. (Neutron diffraction studies, which show the positions of hydrogen atoms, confirm that the carboxylate ion remains as a carboxylate ion throughout and does not actually accept a proton from the imidazole.) Nucleophilic attack by the serine leads to an acylated serine through a tetrahedral intermediate. The new N-terminal end of the cleaved polypeptide chain diffuses away and is replaced by a water molecule.

Regeneration of the active site of chymotrypsin is shown in Fig. 24.20. In this process water acts as the nucleophile and, in a series of steps analogous to those in

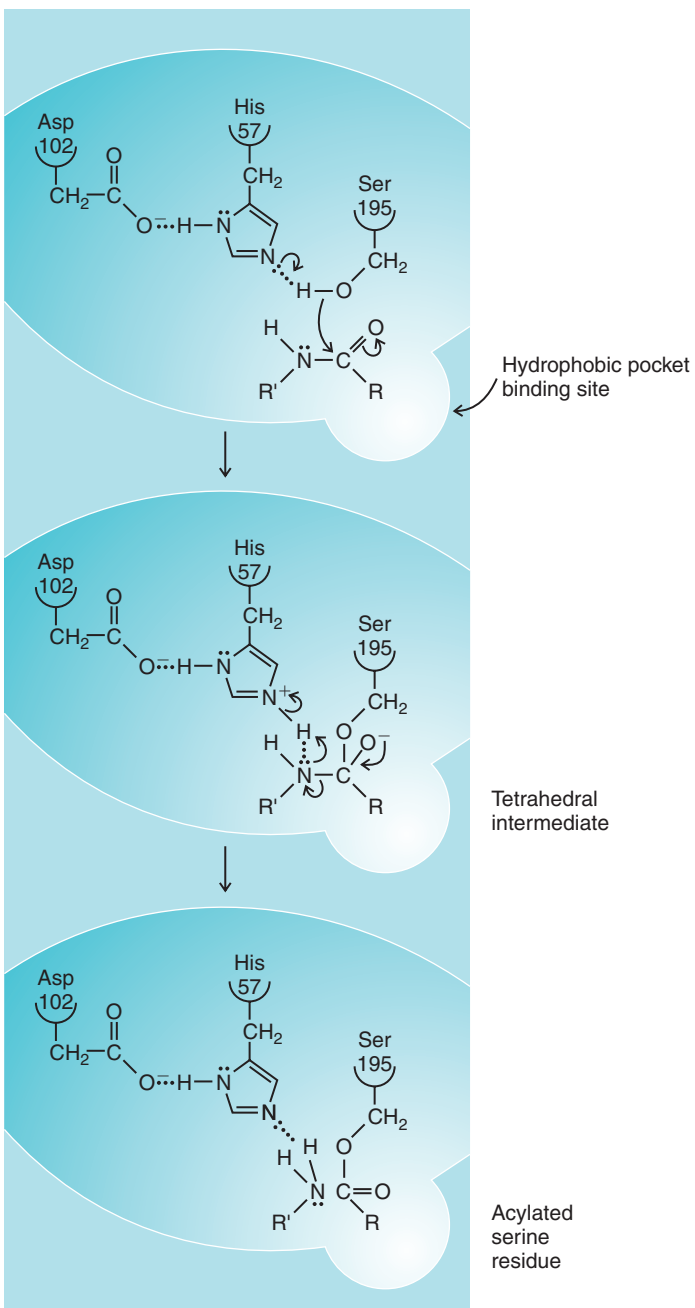


FIGURE 24.19 The catalytic triad of chymotrypsin causes cleavage of a peptide bond by acylation of serine residue 195 of chymotrypsin. Near the active site is a hydrophobic binding site that accommodates nonpolar side chains of the protein.

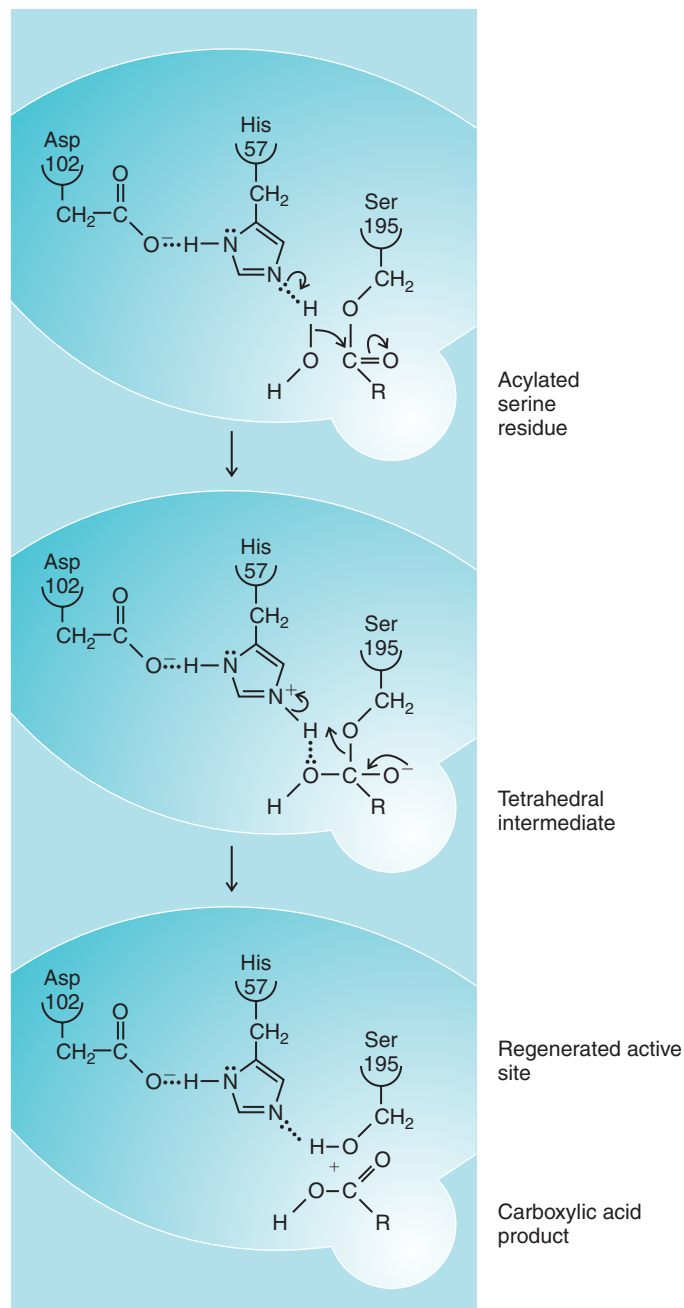
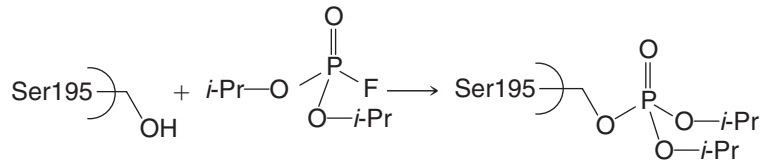


FIGURE 24.20 Regeneration of the active site of chymotrypsin. Water causes hydrolysis of the acyl–serine bond.

Fig. 24.19, hydrolyzes the acyl–serine bond. The enzyme is now ready to repeat the whole process.

There is much evidence for this mechanism that, for reasons of space, we shall have to ignore. One bit of evidence deserves mention, however. There are compounds such as **diisopropylphosphofluoridate (DIPF)** that irreversibly inhibit serine proteases. It has been shown that they do this by reacting only with Ser 195:



Diisopropylphospho-fluoridate (DIPF) **DIP-enzyme**

Recognition of the inactivating effect of DIPF came about as a result of the discovery that DIPF and related compounds are powerful **nerve poisons**. (They are the "nerve gases" of military use, even though they are liquids dispersed as fine droplets, and not gases.) Diisopropylphosphofluoridate inactivates **acetylcholinesterase** (Section 20.3) by reacting with it in the same way that it does with chymotrypsin. Acetylcholinesterase is a **serine esterase** rather than a serine protease.

24.12 HEMOGLOBIN: A CONJUGATED PROTEIN

Some proteins, called **conjugated proteins**, contain as a part of their structure a non-protein group called a **prosthetic group.** An example is the oxygen-carrying protein hemoglobin. Each of the four polypeptide chains of hemoglobin is bound to a prosthetic group called *heme* (Fig. 24.21). The four polypeptide chains of hemoglobin are wound in such a way as to give hemoglobin a roughly spherical shape (Fig. 24.22). Moreover, each heme group lies in a crevice with the hydrophobic vinyl groups of its porphyrin structure surrounded by hydrophobic side chains of amino acid residues. The two propanoate side chains of heme lie near positively charged amino groups of lysine and arginine residues.

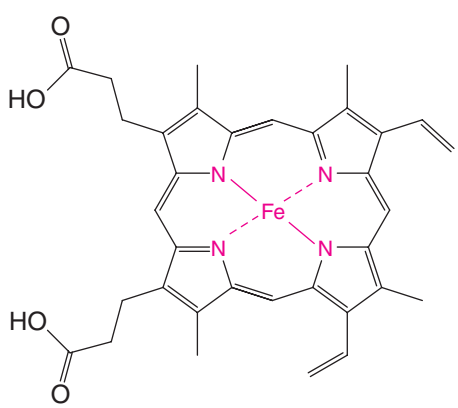


FIGURE 24.21 The structure of heme, the prosthetic group of hemoglobin. Heme has a structure similar to that of chlorophyll (Fig. 22.1) in that each is derived from the heterocyclic ring, porphyrin. The iron of heme is in the ferrous (2+) oxidation state.

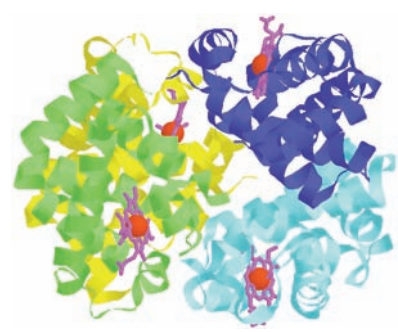


FIGURE 24.22 Hemoglobin. The two α subunits of hemoglobin are shown in blue and green. The two β subunits are shown in yellow and cyan. The four heme groups are shown in purple, and their iron atoms are in red. (PDB ID: IOUU, *http://www.pdb.org*. Tame, J. R., Wilson, J. C., Weber, R. E. The crystal structures of trout Hb I in the deoxy and carbonmonoxy forms. *J. Mol. Biol.* Volume 259, Issue 4, pp. 749–760, 1996.)

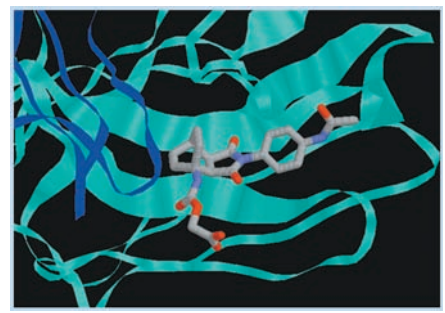
THE CHEMISTRY OF... Some Catalytic Antibodies

Antibodies are chemical warriors of the immune system. Each antibody is a protein produced specifically in response to an invading chemical species (e.g., molecules on the surface of a virus or pollen grain). The purpose of antibodies is to bind with these foreign agents and cause their removal from the organism. The binding of each antibody with its target (the antigen) is usually highly specific.

One way that *catalytic* antibodies have been produced is by prompting an immune response to a chemical species resembling the transition state for a reaction. According to this

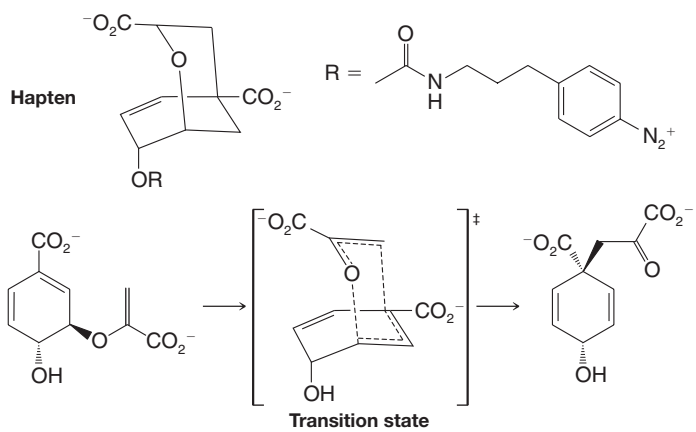
idea, if an antibody is created that preferentially binds with a stable molecule that has a transition state-like structure, other molecules that are capable of reaction *through* this transition state should, in principle, react faster as a result of binding with the antibody. (By facilitating association of the reactants and favoring formation of the transition state structure, the antibody acts in a way similar to an enzyme.) In stunning fashion, precisely this strategy has worked to generate catalytic antibodies for certain Diels–Alder reactions, Claisen rearrangements, and ester hydrolyses. Chemists have synthesized stable molecules that resemble transition states for these reactions, allowed antibodies to be generated against these molecules (called haptens), and then isolated the resulting antibodies. The antibodies thus produced are catalysts when actual substrate molecules are provided.

The following are examples of haptens used as transition state analogs to elicit catalytic antibodies for a Claisen rearrangement, hydrolysis of a carbonate, and a Diels–Alder reaction. The reaction catalyzed by the antibody generated from each hapten is shown as well.

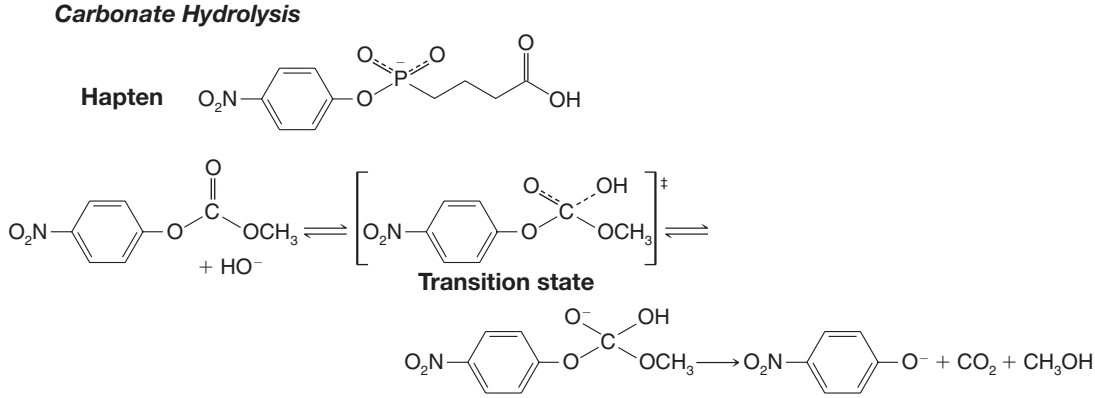


A hapten related to the Diels–Alder adduct from cyclohexadiene and maleimide, bound within a Diels–Alderase catalytic antibody. (PDB ID: 1A4K, http://www.pdb.org. Romesberg, F. E., Spiller, B., Schultz, P. G., Stevens, R. C. Immunological origins of binding and catalysis in a Diels–Alderase antibody. *Science* **279**, pp. 1929–1933, 1998.)

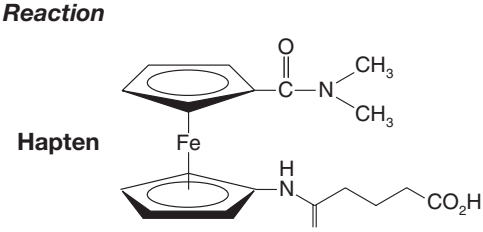
Claisen Rearrangement



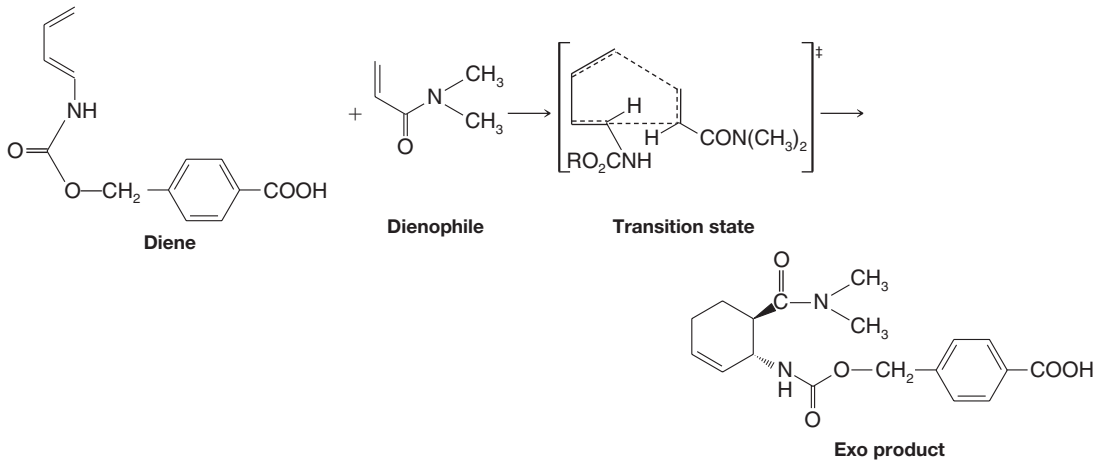
Carbonate Hydrolysis



Diels–Alder Reaction

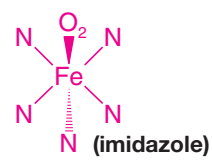


(continues on next page)



This marriage of enzymology and immunology, resulting in chemical offspring, is just one area of exciting research at the interface of chemistry and biology.

The iron of the heme group is in the 2+ (ferrous) oxidation state and it forms a coordinate bond to a nitrogen of the imidazole group of histidine of the polypeptide chain. This leaves one valence of the ferrous ion free to combine with oxygen as follows:



A portion of oxygenated hemoglobin

The fact that the ferrous ion of the heme group combines with oxygen is not particularly remarkable; many similar compounds do the same thing. What is remarkable about hemoglobin is that when the heme combines with oxygen the ferrous ion does not become readily oxidized to the ferric state. Studies with model heme compounds in water, for example, show that they undergo a rapid combination with oxygen but they also undergo a rapid oxidation of the iron from Fe^{2+} to Fe^{3+}. When these same compounds are embedded in the hydrophobic environment of a polystyrene resin, however, the iron is easily oxygenated and deoxygenated, and this occurs *with no change in oxidation state of iron.* In this respect, it is especially interesting to note that X-ray studies of hemoglobin have revealed that the polypeptide chains provide each heme group with a similar hydrophobic environment.

24.13 PURIFICATION AND ANALYSIS OF POLYPEPTIDES AND PROTEINS

24.13A Purification

There are many methods used to purify polypeptides and proteins. The specific methods one chooses depend on the source of the protein (isolation from a natural source or chemical synthesis), its physical properties, including isoelectric point (p*I*), and the quantity of the protein on hand. Initial purification methods may involve precipitation, various forms of column chromatography, and electrophoresis. Perhaps the most important final method for peptide purification, HPLC, is used to purify both peptides generated by automated synthesis and peptides and proteins isolated from nature.

24.13B Analysis

A variety of parameters are used to characterize polypeptides and proteins. One of the most fundamental is molecular weight. Gel electrophoresis can be used to measure the approximate molecular weight of a protein. Gel electrophoresis involves migration of a peptide or protein dissolved in a buffer through a porous polymer gel under the influence

of a high-voltage electric field. The buffer used (typically about pH 9) imparts an overall negative charge to the protein such that the protein migrates toward the positively charged terminal. Migration rate depends on the overall charge and size of the protein as well as the average pore size of the gel. The molecular weight of the protein is inferred by comparing the distance traveled through the gel by the protein of interest with the migration distance of proteins with known molecular weights used as internal standards. The version of this technique called SDS–PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) allows protein molecular weight determinations with an accuracy of about 5–10%.

Mass spectrometry can be used to determine a peptide's molecular weight with very high accuracy and precision. Earlier we discussed mass spectrometry in the context of protein sequencing. Now we shall consider the practical aspects of how molecules with very high molecular weight, such as proteins, can be transferred to the gas phase for mass spectrometric analysis. This is necessary, of course, whether the analysis regards peptide sequencing or full molecular analysis. Small organic molecules, as we discussed in Chapter 9, can be vaporized simply with high vacuum and heat. High-molecular-weight species cannot be transferred to the gas phase solely with heat and vacuum. Fortunately, very effective techniques have been developed for generating gas-phase ions of large molecules without destruction of the sample.

One ionization method is electrospray ionization (ESI, Fig. 24.23), whereby a solution of a peptide (or other analyte) in a volatile solvent containing a trace of acid is sprayed through a high-voltage nozzle into the vacuum chamber of a mass spectrometer. The acid in the solvent generates ions by protonating Lewis basic sites within the analyte. Peptides are typically protonated multiple times. Once injected through the high-voltage nozzle into the vacuum chamber, solvent molecules evaporate from the analyte ions (Fig. 24.23*a*), and the ions are drawn into the mass analyzer (Fig. 24.23*b*). The mass analyzer detects the analyte ions according to their time of flight, and registers their mass-to-charge ratio (*m/z*) (Fig. 24.23*c*). Each peak displayed in the mass spectrum represents the molecular weight of an ion divided by the number of positive charges it carries. From

One-quarter of the 2002 Nobel Prize in Chemistry was awarded to John B. Fenn for his development of ESI mass spectrometry. Another quarter of the prize was awarded to Koichi Tanaku for discoveries that led to matrix-assisted laser desorption ionization (MALDI, see below).

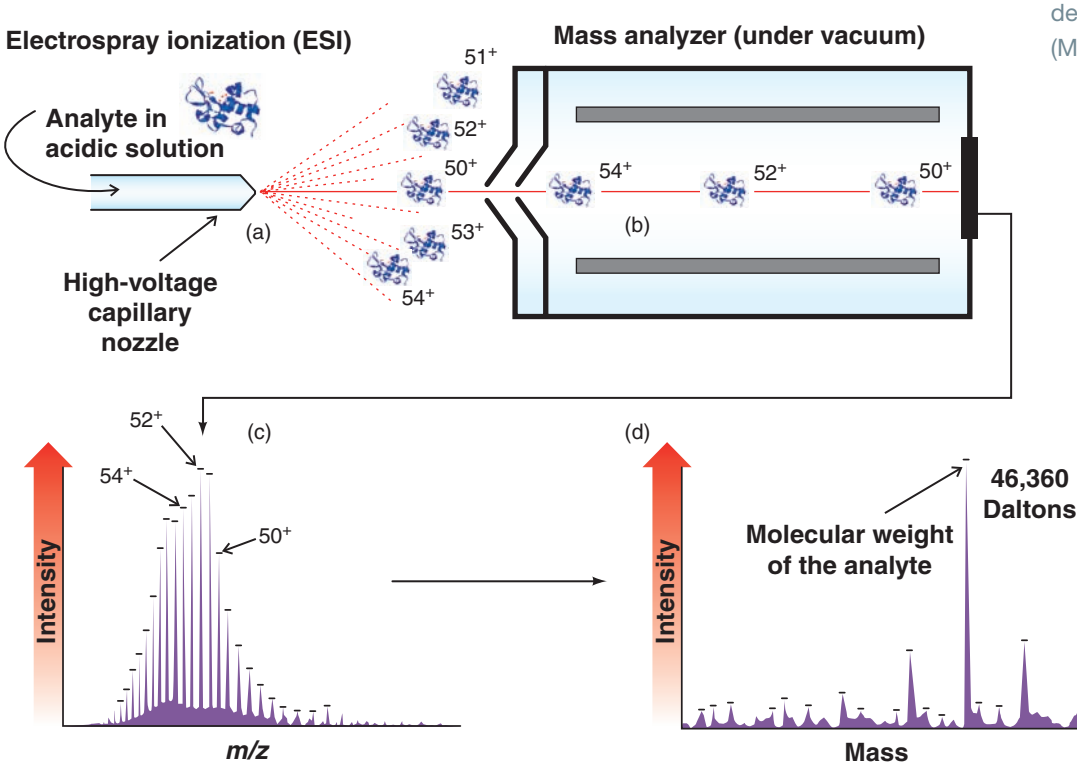


FIGURE 24.23 Electrospray ionization (ESI) mass spectrometry. (*a*) Analyte ions, protonated multiple times by an acidic solvent system, are sprayed through a high-voltage nozzle into a vacuum chamber (diagram is not to scale). Molecules of the solvent evaporate. The mutiply charged analyte ions are drawn into the mass analyzer. (*b*) The analyte ions are separated and detected in the mass analyzer. (*c*) The family of detected ions is displayed in a spectrum according to m/z ratio. (*d*) Computerized deconvolution of the m/z peak series leads to the molecular weight of the analyte.

this series of *m/z* peaks, the molecular weight of the analyte is calculated by a computerized process called deconvolution. An example of a deconvoluted spectrum, indicating a molecular weight of 46,360 atomic mass units (daltons), is shown in Fig. 24.23*d*.

If fragmentation of the analyte molecules is desired, it can be caused by collision-induced dissociation (CID, Section 24.5E). In this case, tandem mass spectrometry is necessary because the first mass analyzer in the system is used to select fragments of the peptide from CID based on their overall mass, while the second mass analyzer in the system records the spectrum of the selected peptide fragment. Multiple fragments from the CID procedure can be analyzed this way. The final spectrum for each peptide fragment selected has the typical appearance of a family of ions, as shown below.

Mass spectrometry with electrospray ionization (ESI-MS) is especially powerful when combined with HPLC because the two techniques can be used in tandem. With such an instrument the effluent from the HPLC is introduced directly into an ESI mass spectrometer. Thus, chromatographic separation of peptides in a mixture and direct structural information about each of them are possible using this technique.

Another method for ionization of nonvolatile molecules is MALDI (matrix-assisted laser desorption ionization, Section 9.18A). Energy from laser bombardment of a sample adsorbed in a solid chemical matrix leads to generation of gas-phase ions that are detected by the mass spectrometer. Both MALDI and ESI are common ionization techniques for the analysis of biopolymers.

24.14 PROTEOMICS

Proteomics and genomics are two fields that have blossomed in recent years. **Proteomics** has to do with the study of all proteins that are expressed in a cell at a given time. **Genomics** (Sections 25.1 and 25.9) focuses on the study of the complete set of genetic instructions in an organism. While the genome holds the instructions for making proteins, it is proteins that carry out the vast majority of functions in living systems. Yet, compared to the tens of thousands of proteins encoded by the genome, we know the structure and function of only a relatively small percentage of proteins in the proteome. For this reason, the field of proteomics has moved to a new level of importance since completion of sequencing the human genome. Many potential developments in health care and medicine now depend on identifying the myriad of proteins that are expressed at any given time in a cell, along with elucidation of their structures and biochemical function. New tools for medical diagnosis and targets for drug design will undoubtedly emerge at an increasing rate as the field of proteomics advances.

One of the basic challenges in proteomics is simply separation of all the proteins present in a cell extract. The next challenge is identification of those proteins that have been separated. Separation of proteins in cell extracts has classically been carried out using two-dimensional polyacrylamide gel electrophoresis (2D PAGE). In 2D PAGE the mixture of proteins extracted from an organism is separated in one dimension of the gel by the isoelectric point (a technique called isoelectric focusing) and in the second dimension by molecular weight. The result is a set of spots in the two-dimensional gel field that represents the location of separated proteins. The protein spots on the gel may then be extracted and analyzed by mass spectrometry or other methods, either as intact proteins or as enzymatic digests. Comparison of the results from mass spectrometry with protein mass spectrometry databases allows identification of many of the proteins separated by the gel.

There are limitations to protein separation by 2D PAGE, however. Not all proteins are amenable to 2D PAGE due to their size, charge, or specific properties. Furthermore, more than one protein may migrate to the same location if their isoelectric points and molecular weights are similar. Finally, 2D PAGE has inherent limits of detection that can leave some proteins of low concentration undetected.

An improvement over 2D PAGE involves two-dimensional microcapillary HPLC coupled with mass spectrometry (see Fig. 24.24). In this technique, called MudPIT (multidimensional protein identification technology, developed by John Yates and co-workers at The Scripps Research Institute), a microcapillary HPLC column is used that has been packed first with a strong cation-exchange resin and then a reversed-phase (hydrophobic) material. The two packing materials used in sequence and with different resolving

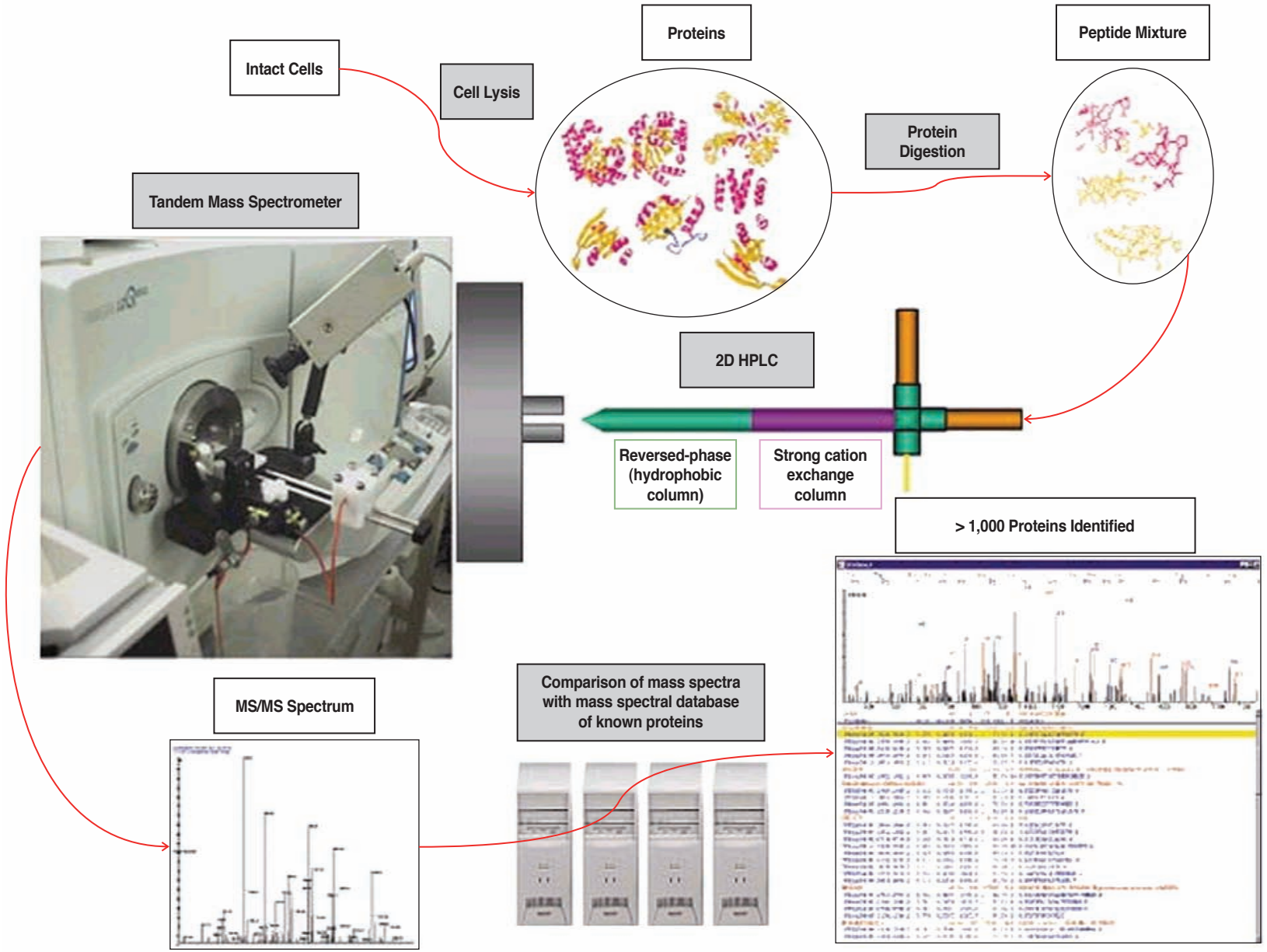


FIGURE 24.24 The high-throughput multidimensional protein identification technology (MudPIT) process. (Reprinted from *International Journal for Parasitology*, December 4, 32(13), Carucci, D.J.; Yates, J. R. 3rd; Florens, L.; Exploring the Proteome of Plasmodium, pp. 1539–1542, Copyright 2002, with permission from Elsevier.)

properties represent the two-dimensional aspect of this technique. A peptide mixture is introduced to the microcapillary column and eluted with pH and solvent gradients over a sequence of automated steps. As the separated peptides are eluted from the column they pass directly into a mass spectrometer. Mass spectrometric data obtained for each protein represent a signature that allows identification of the protein by comparison with a protein mass spectrometry database. This technique of 2D HPLC coupled with mass spectrometry is inherently more sensitive and general than 2D PAGE. One powerful example of its use is the identification by Yates and co-workers of nearly 1500 proteins from the *Saccharomyces cerevisiae* (baker's yeast) proteome in one integrated analysis.

Beyond the identification of proteins, quantitative measurement of the amounts of various proteins that are expressed is also important in proteomics. Various disease states or environmental conditions experienced by a cell may influence the amount of some proteins that are expressed. Quantitative tracking of these changes as a function of cell state could be relevant to studies of disease and the development of therapies. A technique using reagents called isotope-coded affinity tags (ICAT, developed at the University of Washington) allows quantitative analysis and identification of components in complex protein mixtures. The ICAT analysis involves mass spectrometric comparison of isotopically labeled and unlabeled protein segments that have been isolated by affinity chromatography and purified by microcapillary HPLC.

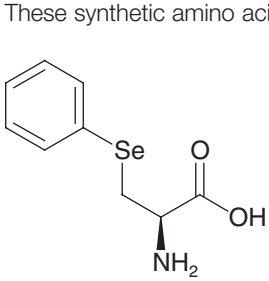
Hand in hand with identification and quantification of proteins remains the need to determine full three-dimensional protein structures. Even though thousands of proteins are encoded in the genome, only a relative handful of them have been studied in depth in terms of detailed structure and function. Full structure determination will therefore continue to be central to the field of proteomics. X-ray crystallography, NMR, and mass spectrometry are key tools that will be applied ever more fervently as the quest intensifies to elucidate as many structures in the proteome as possible.

WHY Do These Topics Matter?]

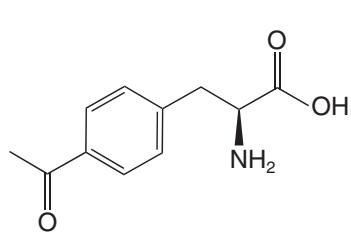
LIFE WITH MORE THAN 20 AMINO ACIDS!?

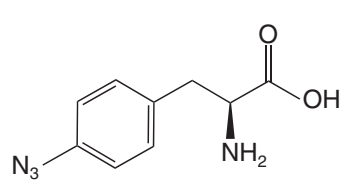
Although the 20 major amino acids that constitute human proteins produce a large and diverse selection of functional molecules, a natural question might be why just 20 amino acids? Why not more? Why not less? And, if 20 is the magic number, then why the specific 20 amino acids that we use? These are very intriguing questions. Although the amino acids in proteins span an array of nonpolar, polar, and fully charged species, arguably many of them are not very different from one another. For instance, leucine and isoleucine differ in the placement of one methyl group, while valine and leucine or serine and threonine or asparagine and glutamine differ by just one carbon atom. Small changes can often lead to profound differences, as we have learned in past chapters, but these acids do not quite seem to be as diverse as they could be. Ultimately, we may never know the answer to these intriguing questions, since they are largely philosophical.

Nevertheless, current knowledge of biochemical systems and synthetic techniques may allow us to explore them from a slightly different perspective. Namely, what would life look like with an expanded genetic code—that is, with additional amino acids added into the proteins of life. Indeed, over the past few years chemists have been able to utilize native biochemical systems as well as evolved tRNA molecules (which we will discuss in Chapter 25) to load many unique amino acids into proteins of interest at any specific point desired in a number of different cells, including those of yeast, some mammals, and bacteria like *E. coli*. Some of the unnatural amino acids are shown below. They include ones with unique metals (like selenium), reactive functional groups (such as a ketone and an azide) that can be used for additional chemistry, and a boronic acid that can be used to bind certain sugars covalently. These synthetic amino acids are all derivatives of phenylalanine, but many other amino acid parent structures can be used as well.

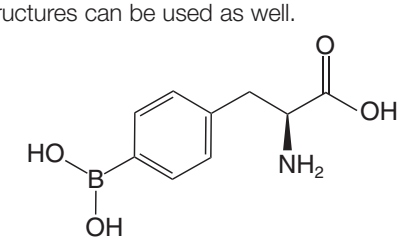


Unique metal; thiophilic





Contains reactive groups to label proteins



Can bind certain carbohydrates

These modifications allow the alteration of an amino acid at the active site of an enzyme, thereby stopping or changing its function. This event makes it possible to determine the potential mechanism of action of the enzyme and/or to observe downstream effects as a result of the alteration. In other cases, a reactive group has allowed chemists to perform labeling experiments and watch what a protein might accomplish in response to an external stimulus. Whether life forms can be generated that have the full machinery to incorporate such additional amino acids on a prolonged basis remains to be seen, but for sure, the ability to load such unnatural amino acids into proteins has led to a wealth of new and valuable knowledge. Indeed, without this capability, in some cases there would be no other way to examine the function of certain proteins.

To learn more about these topics, see:

1. Xie, J.; Schultz, P.G. "Adding amino acids to the genetic repertoire" in *Current Opinion in Chemical Biology* **2005**, *9*, 548–554.
2. Wang, Q; Parrish, A.R.; Wang, L. "Expanding the genetic code for biological studies." *Chem. Biol.* **2009**, *16*, 323–336.
3. Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G. "Expanding the Genetic Code of Escherichia coli". *Science* **2001**, *292*, 498–500.
4. Brustad, E.; Bushey, M. L.; Lee, J. W.; Groff, D.; Liu, W.; Schultz, P. G. "A Genetically Encoded Boronate Amino Acid." *Angew. Chem. Int. Ed.* **2008**, *47*, 8220–8223.

SUMMARY AND REVIEW TOOLS

The study aids for this chapter include key terms and concepts, which are highlighted in bold, blue text within the chapter, defined in the Glossary at the back of the book, and which have hyperlinked definitions in the accompanying *WileyPLUS* course (www.wileyplus.com).

PROBLEMS

Note to Instructors: Many of the homework problems are available for assignment via *WileyPLUS*, an online teaching and learning solution.

STRUCTURE AND REACTIVITY

24.17 **(a)** Which amino acids in Table 24.1 have more than one chirality center?
(b) Write Fischer projections for the isomers of each of these amino acids that would have the L configuration at the α carbon.
(c) What kind of isomers have you drawn in each case?

24.18 **(a)** What product would you expect to obtain from treating tyrosine with excess bromine water?
(b) What product would you expect to be formed in the reaction of phenylalanine with ethanol in the presence of hydrogen chloride?
(c) What product would you expect from the reaction of alanine and benzoyl chloride in aqueous base?

24.19 **(a)** On the basis of the following sequence of reactions, Emil Fischer was able to show that L-(−)-serine and L-(+)-alanine have the same configuration. Write Fischer projections for the intermediates **A–C**:

$$\text{(−)-Serine} \xrightarrow[CH_3OH]{HCl} \mathbf{A}\ (C_4H_{10}ClNO_3) \xrightarrow{PCl_5} \mathbf{B}\ (C_4H_9Cl_2NO_2) \xrightarrow[(2)\ HO^-]{(1)\ H_3O^+,\ H_2O,\ heat} \mathbf{C}\ (C_3H_6ClNO_2) \xrightarrow[\text{dilute } H_3O^+]{Na–Hg} \text{L-(+)-alanine}$$

(b) The configuration of L-(+)-cysteine can be related to that of L-(−)-serine through the following reactions. Write Fischer projections for **D** and **E**:

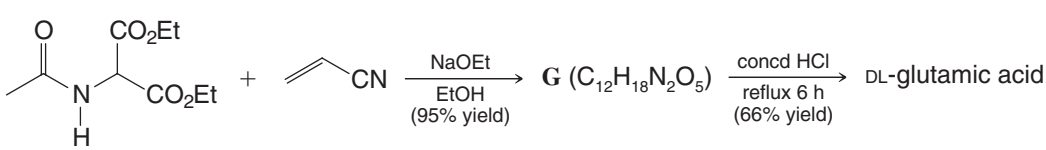
$$\mathbf{B}\ \text{(from part a)} \xrightarrow{HO^-} \mathbf{D}\ (C_4H_8ClNO_2) \xrightarrow{NaSH} \mathbf{E}\ (C_4H_9NO_2S) \xrightarrow[(2)\ HO^-]{(1)\ H_3O^+,\ H_2O,\ heat} \text{L-(+)-cysteine}$$

(c) The configuration of L-(−)-asparagine can be related to that of L-(−)-serine in the following way. What is the structure of **F**?

$$\text{L-(−)-Asparagine} \xrightarrow[\text{Hofmann rearrangement}]{NaOBr/HO^-} \mathbf{F}\ (C_3H_7N_2O_2)$$

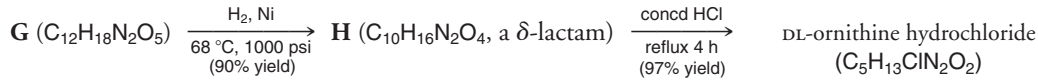
$$\mathbf{C}\ \text{(from part a)} \xrightarrow{NH_3} \mathbf{F}$$

24.20 **(a)** DL-Glutamic acid has been synthesized from diethyl acetamidomalonate in the following way. Outline the reactions involved.



Diethyl acetamidomalonate

(b) Compound **G** has also been used to prepare the amino acid DL-ornithine through the following route. Outline the reactions involved here.



(L-Ornithine is a naturally occurring amino acid but does not occur in proteins. In one metabolic pathway L-ornithine serves as a precursor for L-arginine.)

24.21 Synthetic polyglutamic acid exists as an α helix in solution at pH 2–3. When the pH of such a solution is gradually raised through the addition of a base, a dramatic change in optical rotation takes place at pH 5. This change has been associated with the unfolding of the α helix and the formation of a random coil. What structural feature of polyglutamic acid and what chemical change can you suggest as an explanation of this transformation?

PEPTIDE SEQUENCING

24.22 Bradykinin is a nonapeptide released by blood plasma globulins in response to a wasp sting. It is a very potent pain-causing agent. Its constituent amino acids are 2R, G, 2F, 3P, S. The use of 2,4-dinitrofluorobenzene and carboxypeptidase shows that both terminal residues are arginine. Partial acid hydrolysis of bradykinin gives the following di- and tripeptides:

FS + PGF + PP + SPF + FR + RP

What is the amino acid sequence of bradykinin?

24.23 Complete hydrolysis of a heptapeptide showed that it has the following constituent amino acids:

2A, E, L, K, F, V

Deduce the amino acid sequence of this heptapeptide from the following data.

1. Treatment of the heptapeptide with 2,4-dinitrofluorobenzene followed by incomplete hydrolysis gave, among other products: valine labeled at the α-amino group, lysine labeled at the ε-amino group, and a dipeptide, DNP—VL (DNP = 2,4-dinitrophenyl-).

2. Hydrolysis of the heptapeptide with carboxypeptidase gave an initial high concentration of alanine, followed by a rising concentration of glutamic acid.

3. Partial enzymatic hydrolysis of the heptapeptide gave a dipeptide **(A)** and a tripeptide **(B)**.

(a) Treatment of **A** with 2,4-dinitrofluorobenzene followed by hydrolysis gave DNP-labeled leucine and lysine labeled only at the ε-amino group.

(b) Complete hydrolysis of **B** gave phenylalanine, glutamic acid, and alanine. When **B** was allowed to react with carboxypeptidase, the solution showed an initial high concentration of glutamic acid. Treatment of **B** with 2,4-dinitrofluorobenzene followed by hydrolysis gave labeled phenylalanine.

CHALLENGE PROBLEM

24.24 Part of the evidence for restricted rotation about the carbon–nitrogen bond in a peptide linkage (see Section 24.8A) comes from ^{1}H NMR studies done with simple amides. For example, at room temperature the ^{1}H NMR spectrum of *N,N*-dimethylformamide, $(CH_3)_2NCHO$, shows a doublet at δ 2.80 (3H), a doublet at δ 2.95 (3H), and a multiplet at δ 8.05 (1H). When the spectrum is determined at lower magnetic field strength the doublets are found to have shifted so that the distance (in hertz) that separates one doublet from the other is smaller. When the temperature at which the spectrum is determined is raised, the doublets persist until a temperature of 111 °C is reached; then the doublets coalesce to become a single signal. Explain in detail how these observations are consistent with the existence of a relatively large barrier to rotation about the carbon–nitrogen bond of DMF.

LEARNING GROUP PROBLEMS

1. The enzyme lysozyme and its mechanism are described in Section 24.10. Using the information presented there (and perhaps with additional information from a biochemistry textbook), prepare notes for a class presentation on the mechanism of lysozyme.

2. Chymotrypsin is a member of the serine protease class of enzymes. Its mechanism of action is described in Section 24.11. Using the information presented there (and perhaps supplemented by information from a biochemistry textbook), prepare notes for a class presentation on the mechanism of chymotrypsin. Consider especially the role of the "catalytic triad" with regard to acid–base catalysis and the relative propensity of various groups to act as nucleophiles or leaving groups.