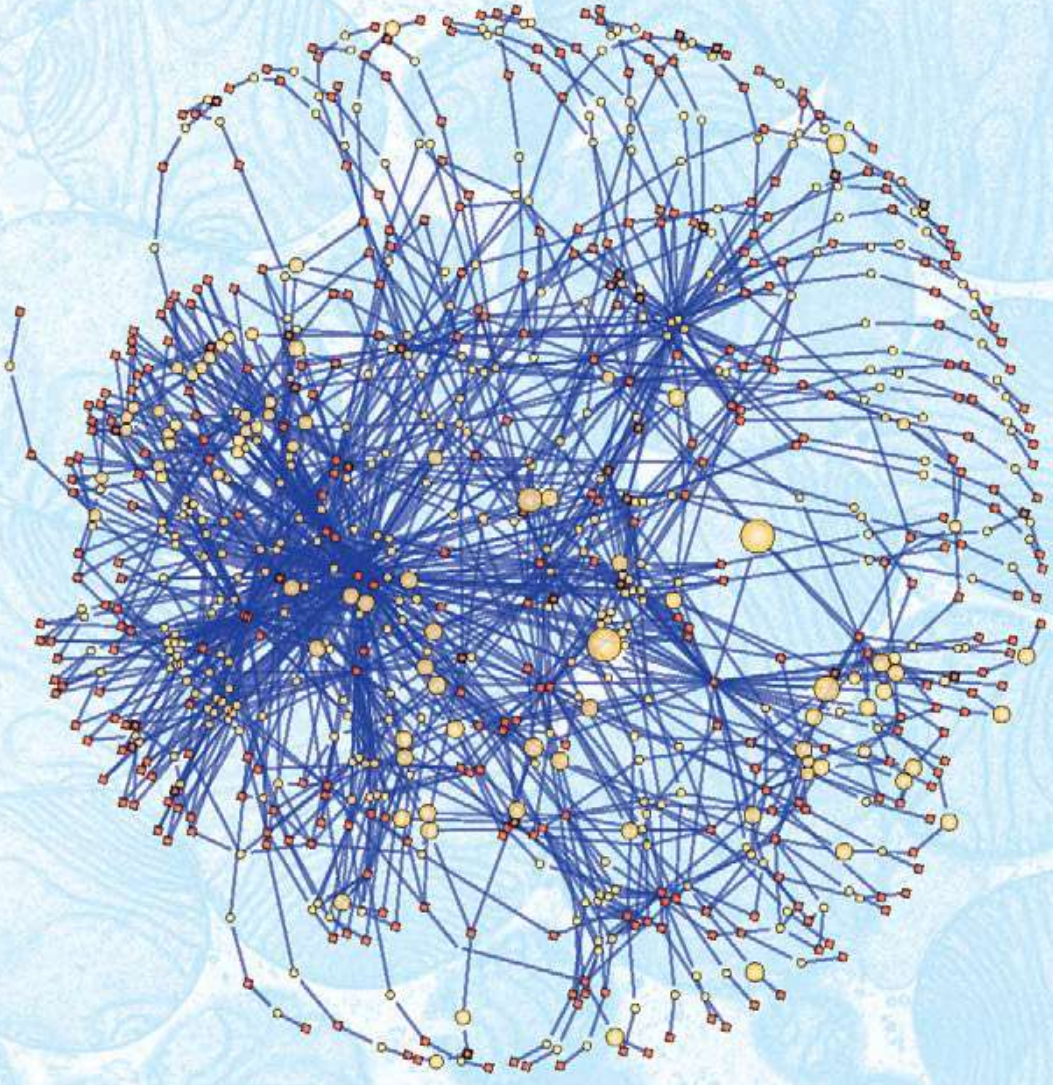
Lehninger

Principles of Biochemistry

SIXTH EDITION

David L. Nelson
Professor of Biochemistry
University of Wisconsin–Madison

Michael M. Cox
Professor of Biochemistry
University of Wisconsin–Madison



W. H. FREEMAN AND COMPANY · New York

Ascendis EX1052
U.S. Patent No. 12,233,106

Publisher:	SUSAN WINSLOW
Senior Acquisitions Editor:	LAUREN SCHULTZ
Senior Developmental Editor:	SUSAN MORAN
Developmental Editor:	MATTHEW TONTONOZ
Associate Director of Marketing:	DEBBIE CLARE
Marketing Director:	JOHN BRITCH
Marketing Assistant	LINDSAY NEFF
Media Editor:	ALLISON MICHAEL
Managing Editor:	PHILIP McCAFFREY
Project Editor:	JANE O'NEILL
Photo Editor:	TED SZCZEPANSKI
Photo Researcher:	ELYSE RIEDER
Art Director:	DIANA BLUME
Illustration Coordinator:	JANICE DONNOLA
Illustrations:	H. ADAM STEINBERG and DRAGONFLY MEDIA GROUP
Molecular Graphics:	H. ADAM STEINBERG
Production Manager:	SUSAN WEIN
Composition:	APTARA, INC.
Printing and binding:	QUAD/GRAPHICS VERSAILLES

North American Edition
Cover image: The network of interactions in an animal mitochondrion. Each dot represents a compound, and each line, an enzyme that interconverts the two compounds. The major nodes include ADP, ATP, NAD^+, and NADH. The image was constructed with Cytoscape software by Anthony Smith in the laboratory of Alan Robinson, Medical Research Council Mitochondrial Biology Unit, Cambridge, UK, using data from MitoMiner (Smith, A.C., Blackshaw, J.A., & Robinson, A.J. (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. *Nucleic Acids Res*. 40, D1160–D1167). *Background*: Transmission electron micrograph of interscapular brown adipose cell from a bat. (Don W. Fawcett/Science Source/Photo Researchers)

International Edition
Cover design: Dirk Kaufman
Cover image: Nastco/iStockphoto.com

Library of Congress Control Number: 2012948755

North American Edition
ISBN-13: 978-1-4292-3414-6
ISBN-10: 1-4292-3414-8

International Edition
ISBN-13: 978-1-4641-0962-1
ISBN-10: 1-4641-0962-1

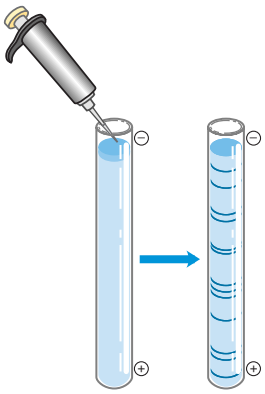
©2013, 2008, 2005, 2000 by W. H. Freeman and Company
All rights reserved

Printed in the United States of America

First printing

W. H. Freeman and Company
41 Madison Avenue
New York, NY 10010
www.whfreeman.com

Macmillan Higher Education
Houndmills, Basingstoke
RG21 6XS, England
www.macmillanhighered.com/international



chapter 3

AMINO ACIDS, PEPTIDES, AND PROTEINS

3.1 Amino Acids 75
3.2 Peptides and Proteins 85
3.3 Working with Proteins 89
3.4 The Covalent Structure of Proteins 96
3.5 Protein Sequences and Evolution 106

The word protein that I propose to you . . . I would wish to derive from *proteios,* because it appears to be the primitive or principal substance of animal nutrition that plants prepare for the herbivores, and which the latter then furnish to the carnivores.

—*J. J. Berzelius,* letter to G. J. Mulder, 1838

Proteins are the most abundant biological macromolecules, occurring in all cells and all parts of cells. Proteins also occur in great variety; thousands of different kinds, ranging in size from relatively small peptides to huge polymers with molecular weights in the millions, may be found in a single cell. Moreover, proteins exhibit enormous diversity of biological function and are the most important final products of the information pathways discussed in Part III of this book. Proteins are the molecular instruments through which genetic information is expressed.

Relatively simple monomeric subunits provide the key to the structure of the thousands of different proteins. All proteins, whether from the most ancient lines of bacteria or from the most complex forms of life, are constructed from the same ubiquitous set of 20 amino acids, covalently linked in characteristic linear sequences. Because each of these amino acids has a side chain with distinctive chemical properties, this group of 20 precursor molecules may be regarded as the alphabet in which the language of protein structure is written.

What is most remarkable is that cells can produce proteins with strikingly different properties and activities by joining the same 20 amino acids in many different combinations and sequences. From these building blocks different organisms can make such widely diverse products as enzymes, hormones, antibodies, transporters, muscle fibers, the lens protein of the eye, feathers, spider webs, rhinoceros horn, milk proteins, antibiotics, mushroom poisons, and myriad other substances having distinct biological activities (Fig. 3–1). Among these protein products, the enzymes are the most varied and specialized. Virtually all cellular reactions are catalyzed by enzymes.

Protein structure and function are the topics of this and the next three chapters. We begin with a description of the fundamental chemical properties of amino acids, peptides, and proteins.

3.1 Amino Acids

Protein Architecture—Amino Acids

Proteins are polymers of amino acids, with each **amino acid residue** joined to its neighbor by a specific type of covalent bond. (The term "residue" reflects the loss of the elements of water when one amino acid is joined to another.) Proteins can be broken down (hydrolyzed) to their constituent amino acids by a variety of methods, and the earliest studies of proteins naturally focused on

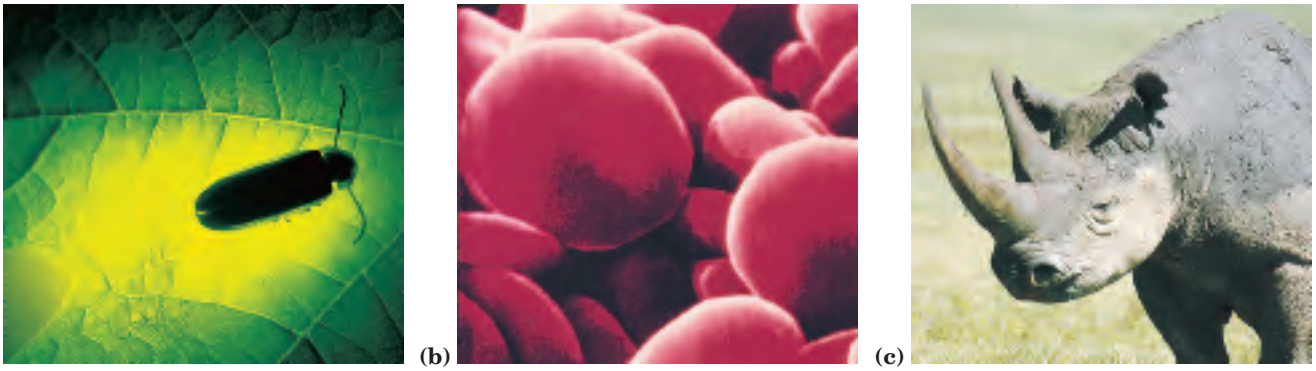


FIGURE 3–1 Some functions of proteins. (a) The light produced by fireflies is the result of a reaction involving the protein luciferin and ATP, catalyzed by the enzyme luciferase (see Box 13–2). **(b)** Erythrocytes contain large amounts of the oxygen-transporting protein hemoglobin. **(c)** The protein keratin, formed by all vertebrates, is the chief structural component of hair, scales, horn, wool, nails, and feathers. The black rhinoceros is nearing extinction in the wild because of the belief prevalent in some parts of the world that a powder derived from its horn has aphrodisiac properties. In reality, the chemical properties of powdered rhinoceros horn are no different from those of powdered bovine hooves or human fingernails.

the free amino acids derived from them. Twenty different amino acids are commonly found in proteins. The first to be discovered was asparagine, in 1806. The last of the 20 to be found, threonine, was not identified until 1938. All the amino acids have trivial or common names, in some cases derived from the source from which they were first isolated. Asparagine was first found in asparagus, and glutamate in wheat gluten; tyrosine was first isolated from cheese (its name is derived from the Greek *tyros,* "cheese"); and glycine (Greek *glykos,* "sweet") was so named because of its sweet taste.

Amino Acids Share Common Structural Features

All 20 of the common amino acids are α-amino acids. They have a carboxyl group and an amino group bonded to the same carbon atom (the α carbon) (Fig. 3–2). They differ from each other in their side chains, or **R groups,** which vary in structure, size, and electric charge, and which influence the solubility of the amino acids in water. In addition to these 20 amino acids there are many less common ones. Some are residues modified after a protein has been synthesized; others are amino acids present in living organisms but not as constituents of proteins. The common amino acids of proteins have been assigned three-letter abbreviations and one-letter symbols (Table 3–1), which are used as shorthand to indicate the composition and sequence of amino acids polymerized in proteins.

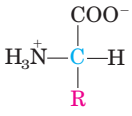
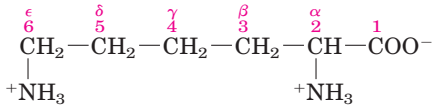


FIGURE 3–2 General structure of an amino acid. This structure is common to all but one of the α-amino acids. (Proline, a cyclic amino acid, is the exception.) The R group or side chain (red) attached to the α carbon (blue) is different in each amino acid.

Two conventions are used to identify the carbons in an amino acid—a practice that can be confusing. The additional carbons in an R group are commonly designated β, γ, δ, ε, and so forth, proceeding out from the α carbon. For most other organic molecules, carbon atoms are simply numbered from one end, giving highest priority (C-1) to the carbon with the substituent containing the atom of highest atomic number. Within this latter convention, the carboxyl carbon of an amino acid would be C-1 and the α carbon would be C-2. In some cases, such as amino acids with heterocyclic R groups, the Greek lettering system is ambiguous and the numbering convention is therefore used.



Lysine

For all the common amino acids except glycine, the α carbon is bonded to four different groups: a carboxyl group, an amino group, an R group, and a hydrogen atom (Fig. 3–2; in glycine, the R group is another hydrogen atom). The α-carbon atom is thus a **chiral center** (p. 17). Because of the tetrahedral arrangement of the bonding orbitals around the α-carbon atom, the four different groups can occupy two unique spatial arrangements, and thus amino acids have two possible stereoisomers. Since they are nonsuperimposable mirror images of each other (Fig. 3–3), the two forms represent a class of stereoisomers called **enantiomers** (see Fig. 1–19). All molecules with a chiral center are also **optically active**—that is, they rotate plane-polarized light (see Box 1–2).

Special nomenclature has been developed to specify the **absolute configuration** of the four substituents of asymmetric carbon atoms. The absolute configurations of simple sugars and amino acids are specified by the **D, L system** (Fig. 3–4), based on the absolute configuration of the three-carbon sugar glyceraldehyde, a convention proposed by Emil Fischer in 1891. (Fischer knew what groups surrounded the asymmetric carbon of glyceraldehyde but had to guess at their absolute configuration; his guess was later confirmed by X-ray diffraction analysis.) For all chiral compounds, stereoisomers having a configuration related to that of L-glyceraldehyde are designated L, and stereoisomers related to D-glyceraldehyde are designated D. The functional groups of L-alanine are matched with those of L-glyceraldehyde by aligning those that can be interconverted by simple, one-step chemical reactions. Thus the carboxyl group of L-alanine occupies the same position about the chiral carbon as does the aldehyde group of L-glyceraldehyde, because an aldehyde is readily converted to a carboxyl group via a one-step oxidation. Historically, the similar *l* and *d* designations were used for levorotatory (rotating light to the left) and dextrorotatory (rotating light to the right). However, not all

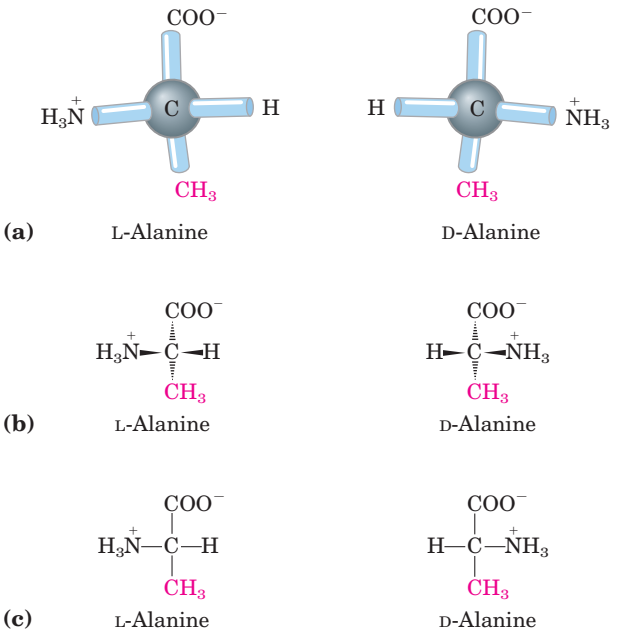


FIGURE 3–3 Stereoisomerism in α-amino acids. (a) The two stereoisomers of alanine, L- and D-alanine, are nonsuperimposable mirror images of each other (enantiomers). **(b, c)** Two different conventions for showing the configurations in space of stereoisomers. In perspective formulas **(b)** the solid wedge-shaped bonds project out of the plane of the paper, the dashed bonds behind it. In projection formulas **(c)** the horizontal bonds are assumed to project out of the plane of the paper, the vertical bonds behind. However, projection formulas are often used casually and are not always intended to portray a specific stereochemical configuration.

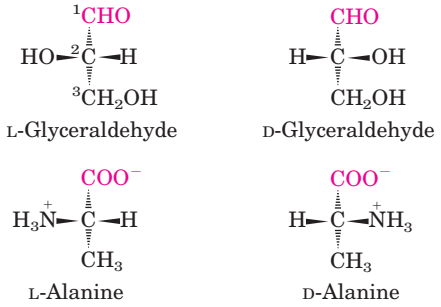


FIGURE 3–4 Steric relationship of the stereoisomers of alanine to the absolute configuration of L- and D-glyceraldehyde. In these perspective formulas, the carbons are lined up vertically, with the chiral atom in the center. The carbons in these molecules are numbered beginning with the terminal aldehyde or carboxyl carbon (red), 1 to 3 from top to bottom as shown. When presented in this way, the R group of the amino acid (in this case the methyl group of alanine) is always below the α carbon. L-Amino acids are those with the α-amino group on the left, and D-amino acids have the α-amino group on the right.

L-amino acids are levorotatory, and the convention shown in Figure 3–4 was needed to avoid potential ambiguities about absolute configuration. By Fischer's convention, L and D refer *only* to the absolute configuration of the four substituents around the chiral carbon, not to optical properties of the molecule.

Another system of specifying configuration around a chiral center is the **RS system,** which is used in the systematic nomenclature of organic chemistry and describes more precisely the configuration of molecules with more than one chiral center (see p. 18).

The Amino Acid Residues in Proteins Are L Stereoisomers

Nearly all biological compounds with a chiral center occur naturally in only one stereoisomeric form, either D or L. The amino acid residues in protein molecules are exclusively L stereoisomers. D-Amino acid residues have been found only in a few, generally small peptides, including some peptides of bacterial cell walls and certain peptide antibiotics.

It is remarkable that virtually all amino acid residues in proteins are L stereoisomers. When chiral compounds are formed by ordinary chemical reactions, the result is a racemic mixture of D and L isomers, which are difficult for a chemist to distinguish and separate. But to a living system, D and L isomers are as different as the right hand and the left. The formation of stable, repeating substructures in proteins (Chapter 4) generally requires that their constituent amino acids be of one stereochemical series. Cells are able to specifically synthesize the L isomers of amino acids because the active sites of enzymes are asymmetric, causing the reactions they catalyze to be stereospecific.

TABLE 3–1 Properties and Conventions Associated with the Common Amino Acids Found in Proteins

Amino acid	Abbreviation/ symbol	M_r	pK_a values: pK_1 (—COOH)	pK_2 ($—NH_3^+$)	pK_R (R group)	pI	Hydropathy index*	Occurrence in proteins (%)†
Nonpolar, aliphatic R groups								
Glycine	Gly G	75	2.34	9.60		5.97	−0.4	7.2
Alanine	Ala A	89	2.34	9.69		6.01	1.8	7.8
Proline	Pro P	115	1.99	10.96		6.48	1.6	5.2
Valine	Val V	117	2.32	9.62		5.97	4.2	6.6
Leucine	Leu L	131	2.36	9.60		5.98	3.8	9.1
Isoleucine	Ile I	131	2.36	9.68		6.02	4.5	5.3
Methionine	Met M	149	2.28	9.21		5.74	1.9	2.3
Aromatic R groups								
Phenylalanine	Phe F	165	1.83	9.13		5.48	2.8	3.9
Tyrosine	Tyr Y	181	2.20	9.11	10.07	5.66	−1.3	3.2
Tryptophan	Trp W	204	2.38	9.39		5.89	−0.9	1.4
Polar, uncharged R groups								
Serine	Ser S	105	2.21	9.15		5.68	−0.8	6.8
Threonine	Thr T	119	2.11	9.62		5.87	−0.7	5.9
Cysteine	Cys C	121	1.96	10.28	8.18	5.07	2.5	1.9
Asparagine	Asn N	132	2.02	8.80		5.41	−3.5	4.3
Glutamine	Gln Q	146	2.17	9.13		5.65	−3.5	4.2
Positively charged R groups								
Lysine	Lys K	146	2.18	8.95	10.53	9.74	−3.9	5.9
Histidine	His H	155	1.82	9.17	6.00	7.59	−3.2	2.3
Arginine	Arg R	174	2.17	9.04	12.48	10.76	−4.5	5.1
Negatively charged R groups								
Aspartate	Asp D	133	1.88	9.60	3.65	2.77	−3.5	5.3
Glutamate	Glu E	147	2.19	9.67	4.25	3.22	−3.5	6.3

*A scale combining hydrophobicity and hydrophilicity of R groups; it can be used to measure the tendency of an amino acid to seek an aqueous environment (− values) or a hydrophobic environment (+ values). See Chapter 11. From Kyte, J. & Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157,** 105–132.

†Average occurrence in more than 1,150 proteins. From Doolittle, R.F. (1989) Redundancies in protein sequences. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G.D., ed.), pp. 599–623, Plenum Press, New York.

Amino Acids Can Be Classified by R Group

Knowledge of the chemical properties of the common amino acids is central to an understanding of biochemistry. The topic can be simplified by grouping the amino acids into five main classes based on the properties of their R groups (Table 3–1), in particular, their **polarity,** or tendency to interact with water at biological pH (near pH 7.0). The polarity of the R groups varies widely, from nonpolar and hydrophobic (water-insoluble) to highly polar and hydrophilic (water-soluble).

The structures of the 20 common amino acids are shown in Figure 3–5, and some of their properties are listed in Table 3–1. Within each class there are gradations of polarity, size, and shape of the R groups.

Nonpolar, Aliphatic R Groups The R groups in this class of amino acids are nonpolar and hydrophobic. The side chains of **alanine, valine, leucine,** and **isoleucine** tend to cluster together within proteins, stabilizing protein structure by means of hydrophobic interactions. **Glycine** has the simplest structure. Although it is formally nonpolar, its very small side chain makes no real contribution to hydrophobic interactions. **Methionine,** one of the two sulfur-containing amino acids, has a nonpolar thioether group in its side chain. **Proline** has an

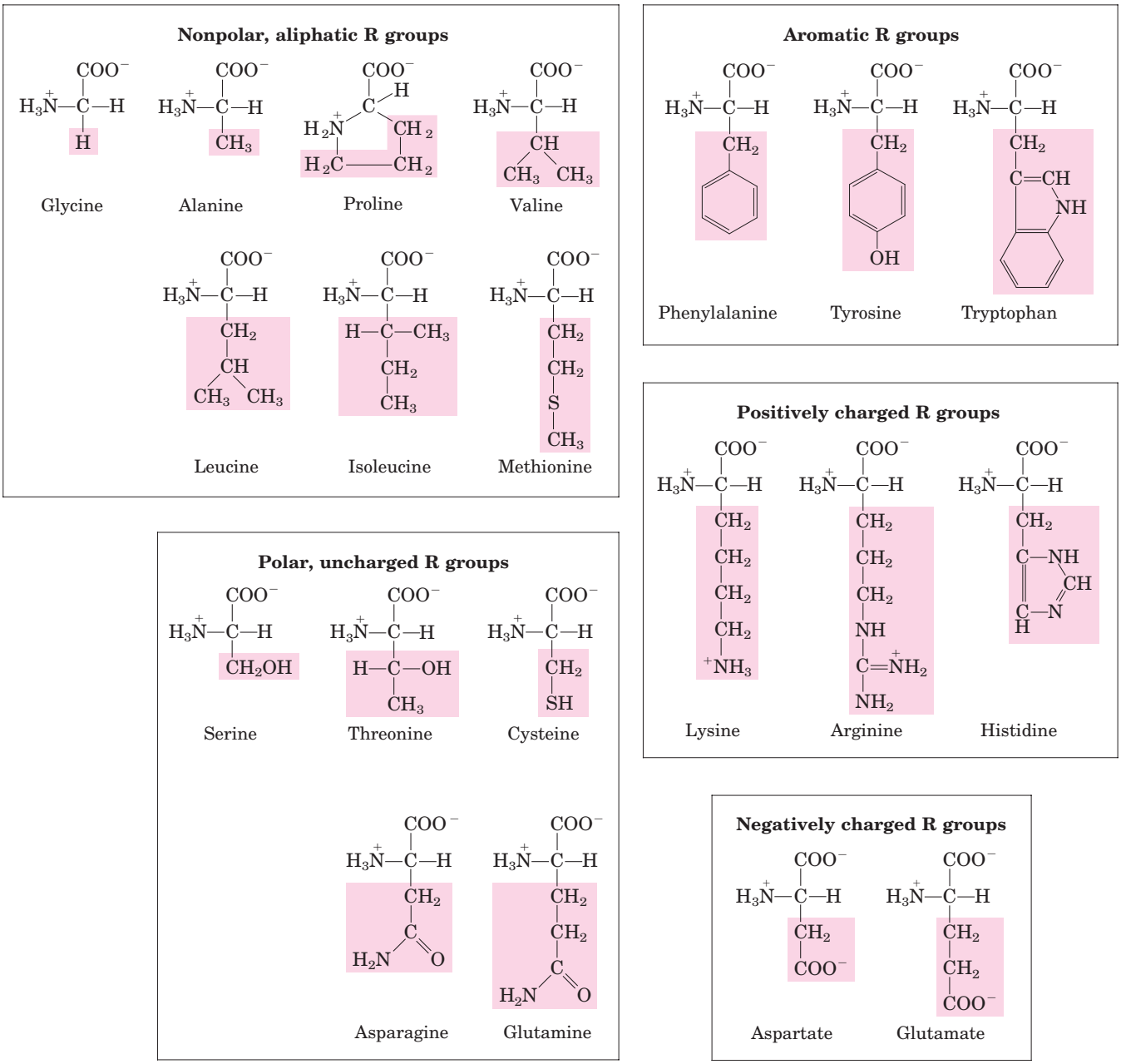


FIGURE 3–5 **The 20 common amino acids of proteins.** The structural formulas show the state of ionization that would predominate at pH 7.0. The unshaded portions are those common to all the amino acids; the portions shaded in red are the R groups. Although the R group of histidine is shown uncharged, its pK_a (see Table 3–1) is such that a small but significant fraction of these groups are positively charged at pH 7.0.

aliphatic side chain with a distinctive cyclic structure. The secondary amino (imino) group of proline residues is held in a rigid conformation that reduces the structural flexibility of polypeptide regions containing proline.

Aromatic R Groups **Phenylalanine, tyrosine,** and **tryptophan,** with their aromatic side chains, are relatively nonpolar (hydrophobic). All can participate in hydrophobic interactions. The hydroxyl group of tyrosine can form hydrogen bonds, and it is an important functional group in some enzymes. Tyrosine and tryptophan are significantly more polar than phenylalanine, because of the tyrosine hydroxyl group and the nitrogen of the tryptophan indole ring.

Tryptophan and tyrosine, and to a much lesser extent phenylalanine, absorb ultraviolet light (Fig. 3–6; Box 3–1). This accounts for the characteristic strong absorbance of light by most proteins at a wavelength of 280 nm, a property exploited by researchers in the characterization of proteins.

Polar, Uncharged R Groups The R groups of these amino acids are more soluble in water, or more hydrophilic, than those of the nonpolar amino acids, because they contain functional groups that form hydrogen bonds with water. This class of amino acids includes **serine, threonine, cysteine, asparagine,** and **glutamine.** The polarity of serine and threonine is contributed by their hydroxyl groups; that of cysteine by its sulfhydryl group; and that of asparagine and glutamine by their amide groups.

Asparagine and glutamine are the amides of two other amino acids also found in proteins, aspartate and glutamate, respectively, to which asparagine and glutamine are easily hydrolyzed by acid or base. Cysteine is readily oxidized to form a covalently linked dimeric amino acid called **cystine,** in which two cysteine molecules or residues are joined by a disulfide bond (Fig. 3–7). The disulfide-linked residues are strongly hydrophobic (nonpolar). Disulfide bonds play a special role in the structures of many proteins by forming covalent links between parts of a protein molecule or between two different polypeptide chains.

Positively Charged (Basic) R Groups The most hydrophilic R groups are those that are either positively or negatively charged. The amino acids in which the R groups have significant positive charge at pH 7.0 are **lysine,** which has a second primary amino group at the ε position on its aliphatic chain; **arginine,** which has a positively charged guanidino group; and **histidine,** which has an imidazole group. Histidine is the only common amino acid having an ionizable side chain with a pK_a near neutrality. In many enzyme-catalyzed reactions, a His residue facilitates the reaction by serving as a proton donor/acceptor.

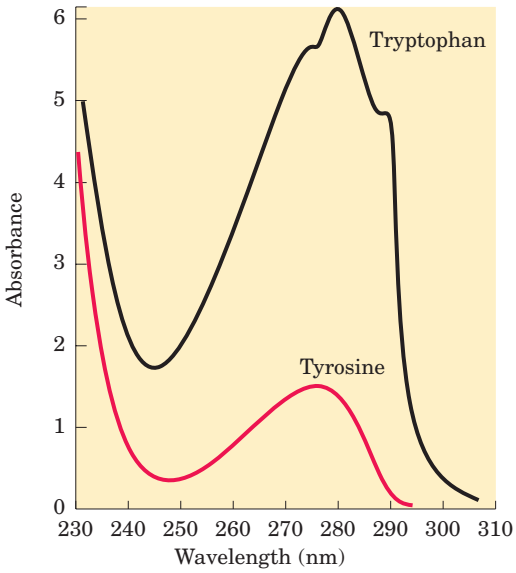


FIGURE 3–6 Absorption of ultraviolet light by aromatic amino acids. Comparison of the light absorption spectra of the aromatic amino acids tryptophan and tyrosine at pH 6.0. The amino acids are present in equimolar amounts (10^{-3} M) under identical conditions. The measured absorbance of tryptophan is as much as four times that of tyrosine. Note that the maximum light absorption for both tryptophan and tyrosine occurs near a wavelength of 280 nm. Light absorption by the third aromatic amino acid, phenylalanine (not shown), generally contributes little to the spectroscopic properties of proteins.

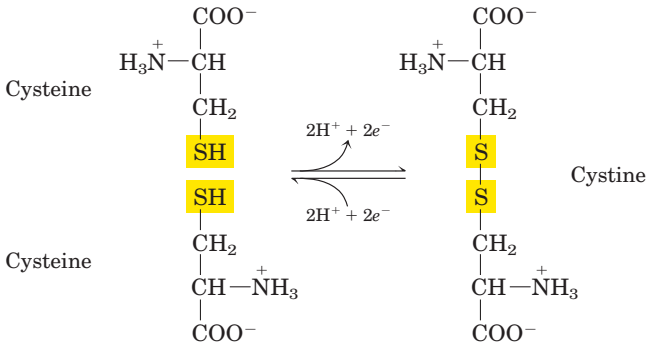


FIGURE 3–7 Reversible formation of a disulfide bond by the oxidation of two molecules of cysteine. Disulfide bonds between Cys residues stabilize the structures of many proteins.

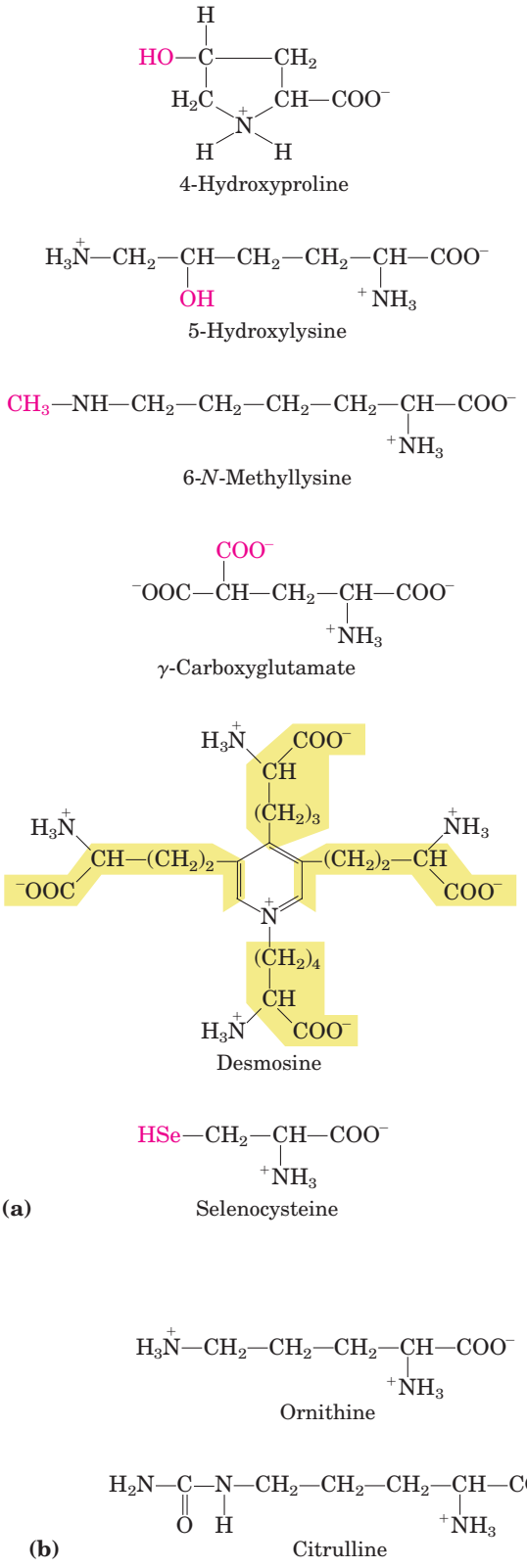
Negatively Charged (Acidic) R Groups The two amino acids having R groups with a net negative charge at pH 7.0 are **aspartate** and **glutamate,** each of which has a second carboxyl group.

Uncommon Amino Acids Also Have Important Functions

In addition to the 20 common amino acids, proteins may contain residues created by modification of common residues already incorporated into a polypeptide (Fig. 3–8a). Among these uncommon amino acids are **4-hydroxyproline,** a derivative of proline, and **5-hydroxylysine,** derived from lysine. The former is found in plant cell wall proteins, and both are found in collagen, a fibrous protein of connective tissues. **6-*N*-Methyllysine** is a constituent of myosin, a contractile protein of muscle. Another important uncommon amino acid is **γ-carboxyglutamate,** found in the blood-clotting protein prothrombin and in certain other proteins that bind Ca^{2+} as part of their biological function. More complex is **desmosine,** a derivative of four Lys residues, which is found in the fibrous protein elastin.

Selenocysteine is a special case. This rare amino acid residue is introduced during protein synthesis rather than created through a postsynthetic modification. It contains selenium rather than the sulfur of cysteine. Actually derived from serine, selenocysteine is a constituent of just a few known proteins.

Some 300 additional amino acids have been found in cells. They have a variety of functions but are not constituents of proteins. **Ornithine** and **citrulline**



(Fig. 3–8b) deserve special note because they are key intermediates (metabolites) in the biosynthesis of arginine (Chapter 22) and in the urea cycle (Chapter 18).

FIGURE 3–8 Uncommon amino acids. (a) Some uncommon amino acids found in proteins. All are derived from common amino acids. Extra functional groups added by modification reactions are shown in red. Desmosine is formed from four Lys residues (the four carbon backbones are shaded in yellow). Note the use of either numbers or Greek letters to identify the carbon atoms in these structures. **(b)** Ornithine and citrulline, which are not found in proteins, are intermediates in the biosynthesis of arginine and in the urea cycle.

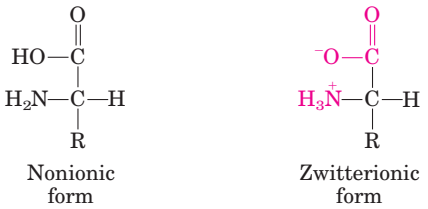
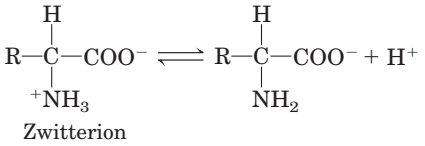


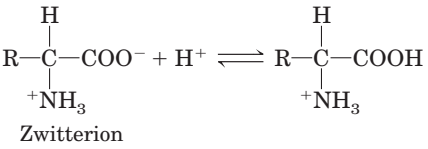
FIGURE 3–9 Nonionic and zwitterionic forms of amino acids. The nonionic form does not occur in significant amounts in aqueous solutions. The zwitterion predominates at neutral pH.

Amino Acids Can Act as Acids and Bases

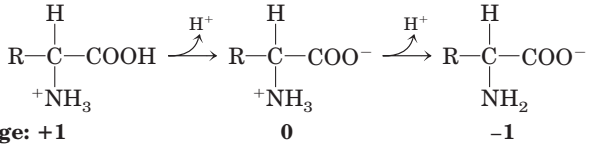
When an amino acid is dissolved in water, it exists in solution as the dipolar ion, or **zwitterion** (German for "hybrid ion"), shown in Figure 3–9. A zwitterion can act as either an acid (proton donor):



or a base (proton acceptor):



Substances having this dual nature are **amphoteric** and are often called **ampholytes** (from "amphoteric electrolytes"). A simple monoamino monocarboxylic α-amino acid, such as alanine, is a diprotic acid when fully protonated—it has two groups, the —COOH group and the $—NH_3^+$ group, that can yield protons:



BOX 3–1 WORKING IN BIOCHEMISTRY

Absorption of Light by Molecules: The Lambert-Beer Law

A wide range of biomolecules absorb light at characteristic wavelengths, just as tryptophan absorbs light at 280 nm (see Fig. 3–6). Measurement of light absorption by a spectrophotometer is used to detect and identify molecules and to measure their concentration in solution. The fraction of the incident light absorbed by a solution at a given wavelength is related to the thickness of the absorbing layer (path length) and the concentration of the absorbing species (Fig. 1). These two relationships are combined into the Lambert-Beer law,

$$\log \frac{I_0}{I} = \varepsilon c l$$

where I_0 is the intensity of the incident light, I is the intensity of the transmitted light, ε is the molar extinction coefficient (in units of liters per mole-centimeter), c is the concentration of the absorbing species (in moles per liter), and l is the path length of the light-absorbing sample (in centimeters). The Lambert-Beer law assumes that the incident light is parallel and monochromatic (of a single wavelength) and that the solvent and solute molecules are randomly oriented. The expression $\log (I_0/I)$ is called the **absorbance,** designated A.

It is important to note that each successive millimeter of path length of absorbing solution in a 1.0 cm cell absorbs not a constant amount but a constant fraction of the light that is incident upon it. However, with an absorbing layer of fixed path length, *the absorbance,* A, *is directly proportional to the concentration of the absorbing solute.*

The molar extinction coefficient varies with the nature of the absorbing compound, the solvent, and the wavelength, and also with pH if the light-absorbing species is in equilibrium with an ionization state that has different absorbance properties.

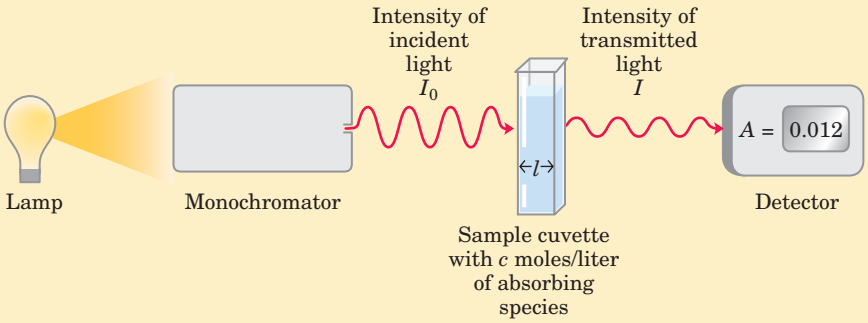


FIGURE 1 The principal components of a spectrophotometer. A light source emits light along a broad spectrum, then the monochromator selects and transmits light of a particular wavelength. The monochromatic light passes through the sample in a cuvette of path length l and is absorbed by the sample in proportion to the concentration of the absorbing species. The transmitted light is measured by a detector.

Amino Acids Have Characteristic Titration Curves

Acid-base titration involves the gradual addition or removal of protons (Chapter 2). Figure 3–10 shows the titration curve of the diprotic form of glycine. The plot has two distinct stages, corresponding to deprotonation of two different groups on glycine. Each of the two stages resembles in shape the titration curve of a monoprotic acid, such as acetic acid (see Fig. 2–17), and can be analyzed in the same way. At very low pH, the predominant ionic species of glycine is the fully protonated form, $^{+}H_3N—CH_2—COOH$. At the midpoint in the first stage of the titration, in which the —COOH group of glycine loses its proton, equimolar concentrations of the proton-donor ($^{+}H_3N—CH_2—COOH$) and proton-acceptor ($^{+}H_3N—CH_2—COO^{-}$) species are present. At the midpoint of any titration, a point of inflection is reached where the pH is equal to the pK_a of the protonated group being titrated (see Fig. 2–18). For glycine, the pH at the midpoint is 2.34, thus its —COOH group has a pK_a (labeled pK_1 in Fig. 3–10) of 2.34. (Recall from Chapter 2 that pH and pK_a are simply convenient notations for proton concentration and the equilibrium constant for ionization, respectively. The pK_a is a measure of the tendency of a group to give up a proton, with that tendency decreasing tenfold as the pK_a increases by one unit.) As the titration proceeds, another important point is reached at pH 5.97. Here there is another point of inflection, at which removal of the first proton is essentially complete and removal of the second has just begun. At this pH glycine is present largely as the dipolar ion $^{+}H_3N—CH_2—COO^{-}$. We shall return to the significance of this inflection point in the titration curve (labeled pI in Fig. 3–10) shortly.

The second stage of the titration corresponds to the removal of a proton from the $—NH_3^{+}$ group of glycine. The pH at the midpoint of this stage is 9.60, equal to the pK_a (labeled pK_2 in Fig. 3–10) for the $—NH_3^{+}$ group. The titration is essentially complete at a pH of about 12, at which point the predominant form of glycine is $H_2N—CH_2—COO^{-}$.

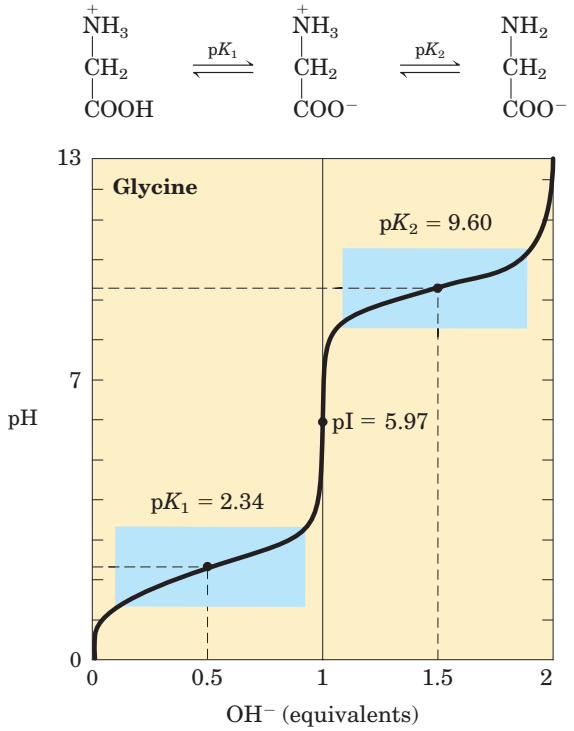


FIGURE 3–10 **Titration of an amino acid.** Shown here is the titration curve of 0.1 M glycine at 25 °C. The ionic species predominating at key points in the titration are shown above the graph. The shaded boxes, centered at about $pK_1 = 2.34$ and $pK_2 = 9.60$, indicate the regions of greatest buffering power.

From the titration curve of glycine we can derive several important pieces of information. First, it gives a quantitative measure of the pK_a of each of the two ionizing groups: 2.34 for the —COOH group and 9.60 for the $—NH_3^+$ group. Note that the carboxyl group of glycine is over 100 times more acidic (more easily ionized) than the carboxyl group of acetic acid, which, as we saw in Chapter 2, has a pK_a of 4.76—about average for a carboxyl group attached to an otherwise unsubstituted aliphatic hydrocarbon. The perturbed pK_a of glycine is caused by repulsion between the departing proton and the nearby positively charged amino group on the α-carbon atom, as described in Figure 3–11. The opposite charges on the resulting zwitterion are stabilizing, nudging the equilibrium farther to the right. Similarly, the pK_a of the amino group in glycine is perturbed downward relative to the average pK_a of an amino group. This effect is due partly to the electronegative oxygen atoms in the carboxyl groups, which tend to pull electrons toward them, increasing the tendency of the amino group to give up a proton. Hence, the α-amino group has a pK_a that is lower than that of an aliphatic amine such as methylamine (Fig. 3–11). In short, the pK_a of any functional group is greatly affected by its chemical environment, a phenomenon sometimes exploited in the active sites of enzymes to promote exquisitely adapted reaction mechanisms that depend on the perturbed pK_a values of proton donor/acceptor groups of specific residues.

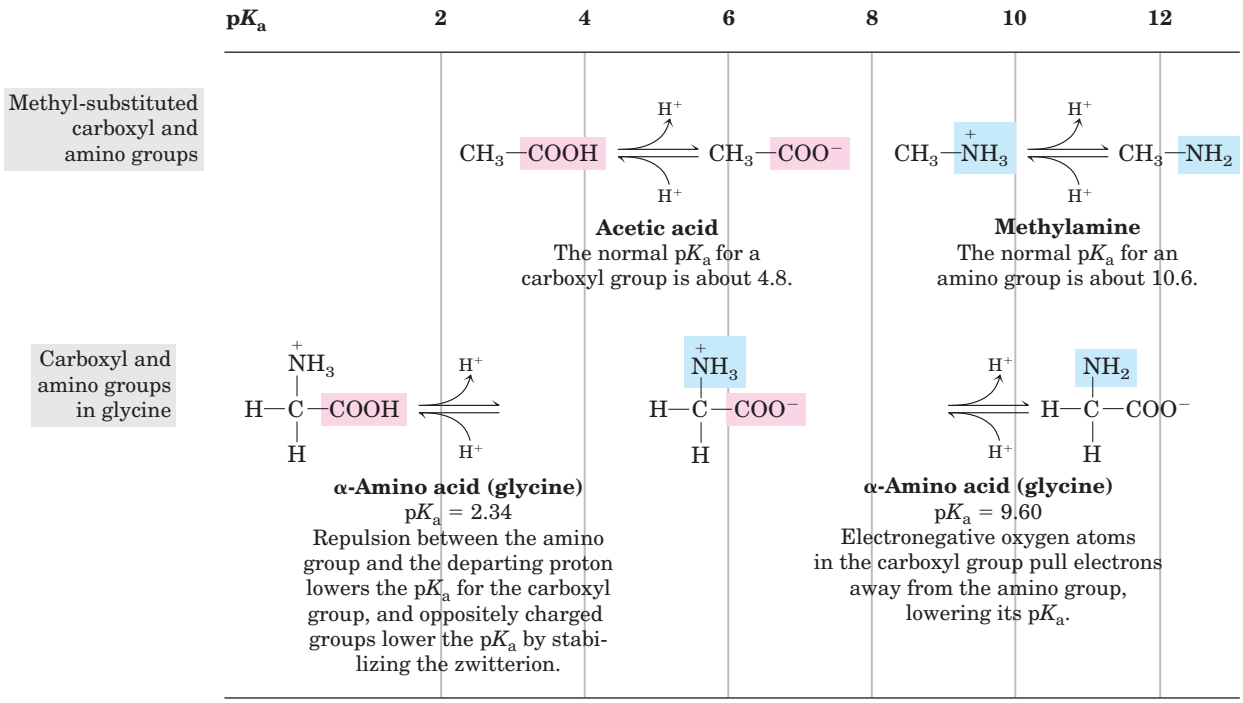


FIGURE 3–11 **Effect of the chemical environment on pK_a.** The pK_a values for the ionizable groups in glycine are lower than those for simple, methyl-substituted amino and carboxyl groups. These downward perturbations of pK_a are due to intramolecular interactions. Similar effects can be caused by chemical groups that happen to be positioned nearby—for example, in the active site of an enzyme.

The second piece of information provided by the titration curve of glycine is that this amino acid has *two* regions of buffering power. One of these is the relatively flat portion of the curve, extending for approximately 1 pH unit on either side of the first pK_a of 2.34, indicating that glycine is a good buffer near this pH. The other buffering zone is centered around pH 9.60. (Note that glycine is not a good buffer at the pH of intracellular fluid or blood, about 7.4.) Within the buffering ranges of glycine, the Henderson-Hasselbalch equation (see Box 2–3) can be used to calculate the proportions of proton-donor and proton-acceptor species of glycine required to make a buffer at a given pH.

Titration Curves Predict the Electric Charge of Amino Acids

Another important piece of information derived from the titration curve of an amino acid is the relationship between its net electric charge and the pH of the solution. At pH 5.97, the point of inflection between the two stages in its titration curve, glycine is present predominantly as its dipolar form, fully ionized but with no *net* electric charge (Fig. 3–10). The characteristic pH at which the net electric charge is zero is called the **isoelectric point** or **isoelectric pH,** designated **pI.** For glycine, which has no ionizable group in its side chain, the isoelectric point is simply the arithmetic mean of the two pK_a values:

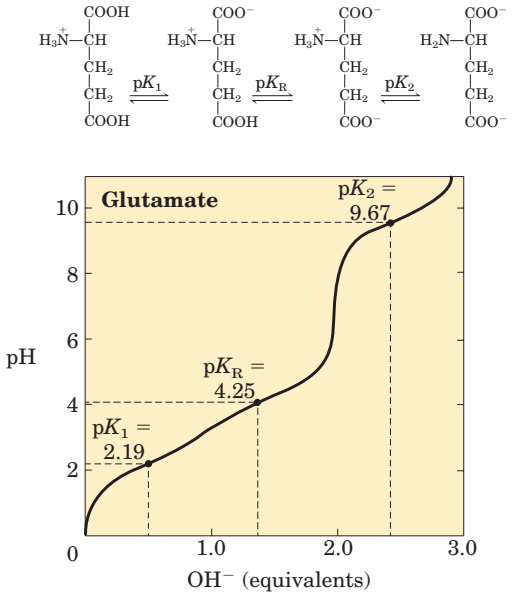
$$\text{pI} = \frac{1}{2}(\text{p}K_1 + \text{p}K_2) = \frac{1}{2}(2.34 + 9.60) = 5.97$$

As is evident in Figure 3–10, glycine has a net negative charge at any pH above its pI and will thus move toward the positive electrode (the anode) when placed in an electric field. At any pH below its pI, glycine has a net positive charge and will move toward the negative electrode (the cathode). The farther the pH of a glycine solution is from its isoelectric point, the greater the net electric charge of the population of glycine molecules. At pH 1.0, for example, glycine exists almost entirely as the form $^{+}H_3N—CH_2—COOH$, with a net positive charge of 1.0. At pH 2.34, where there is an equal mixture of $^{+}H_3N—CH_2—COOH$ and $^{+}H_3N—CH_2—COO^{-}$, the average or net positive charge is 0.5. The sign and the magnitude of the net charge of any amino acid at any pH can be predicted in the same way.

Amino Acids Differ in Their Acid-Base Properties

The shared properties of many amino acids permit some simplifying generalizations about their acid-base behaviors. First, all amino acids with a single α-amino group, a single α-carboxyl group, and an R group that does not ionize have titration curves resembling that of glycine (Fig. 3–10). These amino acids have very similar, although not identical, pK_a values: pK_a of the —COOH group in the range of 1.8 to 2.4, and pK_a of the $—NH_3^+$ group in the range of 8.8 to 11.0 (Table 3–1).

Second, amino acids with an ionizable R group have more complex titration curves, with *three* stages corresponding to the three possible ionization steps; thus they have three pK_a values. The additional stage for the titration of the ionizable R group merges to some extent with the other two. The titration curves for two amino acids of this type, glutamate and histidine, are shown in Figure 3–12. The isoelectric points reflect the nature of the ionizing R groups present. For example, glutamate



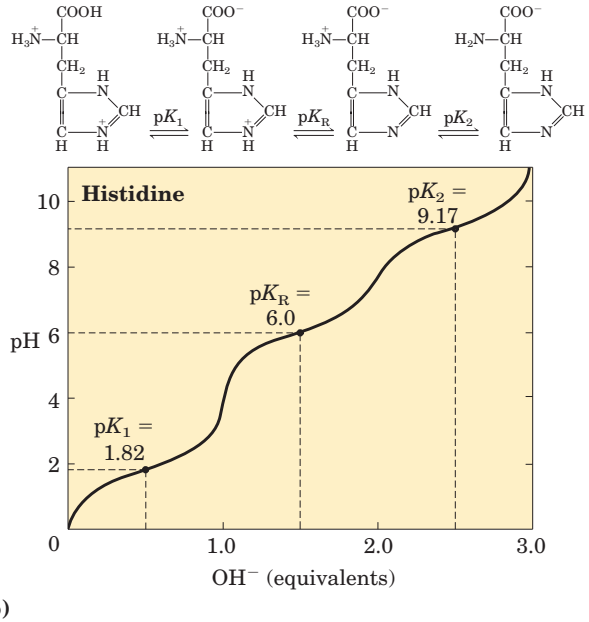


FIGURE 3–12 Titration curves for **(a)** glutamate and **(b)** histidine. The pK_a of the R group is designated here as pK_R.

has a pI of 3.22, considerably lower than that of glycine. This is due to the presence of two carboxyl groups, which, at the average of their pK_a values (3.22), contribute a net charge of −1 that balances the +1 contributed by the amino group. Similarly, the pI of histidine, with two groups that are positively charged when protonated, is 7.59 (the average of the pK_a values of the amino and imidazole groups), much higher than that of glycine.

Finally, as pointed out earlier, under the general condition of free and open exposure to the aqueous environment, only histidine has an R group (pK_a = 6.0) providing significant buffering power near the neutral pH usually found in the intracellular and extracellular fluids of most animals and bacteria (Table 3–1).

SUMMARY 3.1 Amino Acids

- The 20 amino acids commonly found as residues in proteins contain an α-carboxyl group, an α-amino group, and a distinctive R group substituted on the α-carbon atom. The α-carbon atom of all amino acids except glycine is asymmetric, and thus amino acids can exist in at least two stereoisomeric forms. Only the L stereoisomers, with a configuration related to the absolute configuration of the reference molecule L-glyceraldehyde, are found in proteins.
- Other, less common amino acids also occur, either as constituents of proteins (through modification of common amino acid residues after protein synthesis) or as free metabolites.
- Amino acids are classified into five types on the basis of the polarity and charge (at pH 7) of their R groups.
- Amino acids vary in their acid-base properties and have characteristic titration curves. Monoamino monocarboxylic amino acids (with nonionizable R groups) are diprotic acids ($^+H_3NCH(R)COOH$) at low pH and exist in several different ionic forms as the pH is increased. Amino acids with ionizable R groups have additional ionic species, depending on the pH of the medium and the pK_a of the R group.

3.2 Peptides and Proteins

We now turn to polymers of amino acids, the **peptides** and **proteins.** Biologically occurring polypeptides range in size from small to very large, consisting of two or three to thousands of linked amino acid residues. Our focus is on the fundamental chemical properties of these polymers.

Peptides Are Chains of Amino Acids

Two amino acid molecules can be covalently joined through a substituted amide linkage, termed a **peptide bond,** to yield a dipeptide. Such a linkage is formed by removal of the elements of water (dehydration) from the α-carboxyl group of one amino acid and the α-amino group of another (Fig. 3–13). Peptide bond formation is an example of a condensation reaction, a common class of reactions in living cells. Under standard biochemical conditions, the equilibrium for the reaction shown in Figure 3–13 favors the amino acids over the dipeptide. To make the reaction thermodynamically more favorable, the carboxyl group must be chemically modified or activated so that the hydroxyl group can be more readily eliminated. A chemical approach to this problem is outlined later in this chapter. The biological approach to peptide bond formation is a major topic of Chapter 27.

Three amino acids can be joined by two peptide bonds to form a tripeptide; similarly, amino acids can be linked to form tetrapeptides, pentapeptides, and so forth. When a few amino acids are joined in this fashion, the structure is called an **oligopeptide.** When many amino acids are joined, the product is called a **polypeptide.** Proteins may have thousands of amino acid residues. Although the terms "protein" and "polypeptide" are sometimes used interchangeably, molecules referred to as polypeptides generally have molecular weights below 10,000, and those called proteins have higher molecular weights.

Figure 3–14 shows the structure of a pentapeptide. As already noted, an amino acid unit in a peptide is often called a residue (the part left over after losing a hydrogen atom from its amino group and the hydroxyl moiety from its carboxyl group). In a peptide, the amino acid residue at the end with a free α-amino group is the **amino-terminal** (or *N*-terminal) residue; the residue

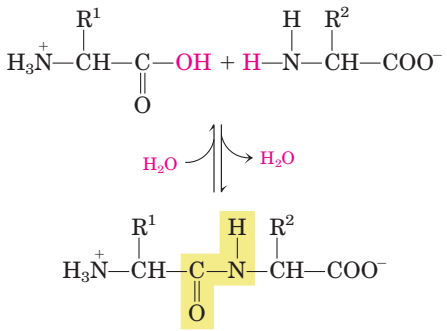


FIGURE 3–13 Formation of a peptide bond by condensation. The α-amino group of one amino acid (with R^2 group) acts as a nucleophile to displace the hydroxyl group of another amino acid (with R^1 group), forming a peptide bond (shaded in yellow). Amino groups are good nucleophiles, but the hydroxyl group is a poor leaving group and is not readily displaced. At physiological pH, the reaction shown does not occur to any appreciable extent.

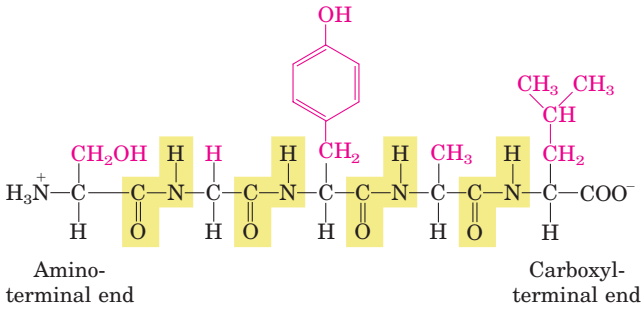


FIGURE 3–14 The pentapeptide serylglycyltyrosylalanylleucine, or Ser–Gly–Tyr–Ala–Leu. Peptides are named beginning with the amino-terminal residue, which by convention is placed at the left. The peptide bonds are shaded in yellow; the R groups are in red.

at the other end, which has a free carboxyl group, is the **carboxyl-terminal** (*C*-terminal) residue.

Although hydrolysis of a peptide bond is an exergonic reaction, it occurs slowly because of its high activation energy. As a result, the peptide bonds in proteins are quite stable, with an average half-life ($t_{1/2}$) of about 7 years under most intracellular conditions.

Peptides Can Be Distinguished by Their Ionization Behavior

Peptides contain only one free α-amino group and one free α-carboxyl group, at opposite ends of the chain (Fig. 3–15). These groups ionize as they do in free amino acids, although the ionization constants are different because an oppositely charged group is no longer linked to the α carbon. The α-amino and α-carboxyl groups of all nonterminal amino acids are covalently joined in the peptide bonds, which do not ionize and thus do not contribute to the total acid-base behavior of peptides. However, the R groups of some amino acids can ionize (Table 3–1), and in a peptide these contribute to the overall acid-base properties of the molecule (Fig. 3–15). Thus the acid-base behavior of a peptide can be predicted from its free α-amino and α-carboxyl groups as well as the nature and number of its ionizable R groups.

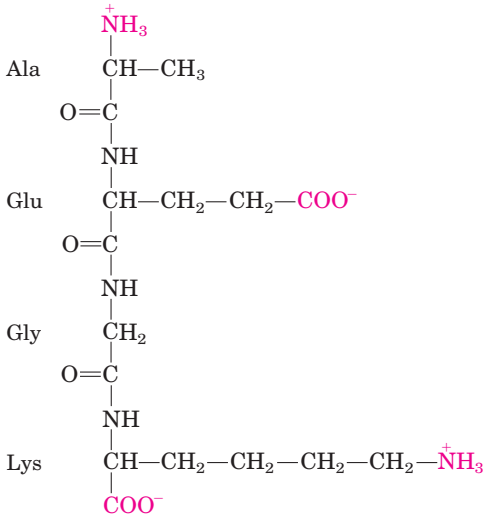
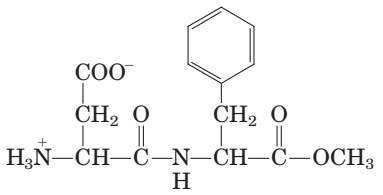


FIGURE 3–15 **Alanylglutamylglycyllysine.** This tetrapeptide has one free α-amino group, one free α-carboxyl group, and two ionizable R groups. The groups ionized at pH 7.0 are in red.

Like free amino acids, peptides have characteristic titration curves and a characteristic isoelectric pH (pI) at which they do not move in an electric field. These properties are exploited in some of the techniques used to separate peptides and proteins, as we shall see later in the chapter. It should be emphasized that the pK_a value for an ionizable R group can change somewhat when an amino acid becomes a residue in a peptide. The loss of charge in the α-carboxyl and α-amino groups, the interactions with other peptide R groups, and other environmental factors can affect the pK_a. The pK_a values for R groups listed in Table 3–1 can be a useful guide to the pH range in which a given group will ionize, but they cannot be strictly applied to peptides.

Biologically Active Peptides and Polypeptides Occur in a Vast Range of Sizes

No generalizations can be made about the molecular weights of biologically active peptides and proteins in relation to their functions. Naturally occurring peptides range in length from two to many thousands of amino acid residues. Even the smallest peptides can have biologically important effects. Consider the commercially synthesized dipeptide L-aspartyl-L-phenylalanine methyl ester, the artificial sweetener better known as aspartame or NutraSweet.



L-Aspartyl-L-phenylalanine methyl ester (aspartame)

Many small peptides exert their effects at very low concentrations. For example, a number of vertebrate hormones (Chapter 23) are small peptides. These include oxytocin (nine amino acid residues), which is secreted by the posterior pituitary and stimulates uterine contractions; bradykinin (nine residues), which inhibits inflammation of tissues; and thyrotropin-releasing factor (three residues), which is formed in the hypothalamus and stimulates the release of another hormone, thyrotropin, from the anterior pituitary gland. Some extremely toxic mushroom poisons, such as amanitin, are also small peptides, as are many antibiotics.

Slightly larger are small polypeptides and oligopeptides such as the pancreatic hormone insulin, which contains two polypeptide chains, one having 30 amino acid

residues and the other 21. Glucagon, another pancreatic hormone, has 29 residues; it opposes the action of insulin. Corticotropin is a 39-residue hormone of the anterior pituitary gland that stimulates the adrenal cortex.

How long are the polypeptide chains in proteins? As Table 3–2 shows, lengths vary considerably. Human cytochrome *c* has 104 amino acid residues linked in a single chain; bovine chymotrypsinogen has 245 residues. At the extreme is titin, a constituent of vertebrate muscle, which has nearly 27,000 amino acid residues and a molecular weight of about 3,000,000. The vast majority of naturally occurring proteins are much smaller than this, containing fewer than 2,000 amino acid residues.

Some proteins consist of a single polypeptide chain, but others, called **multisubunit** proteins, have two or more polypeptides associated noncovalently (Table 3–2). The individual polypeptide chains in a multisubunit protein may be identical or different. If at least two are identical the protein is said to be **oligomeric,** and the identical units (consisting of one or more polypeptide chains) are referred to as **protomers.** Hemoglobin, for example, has four polypeptide subunits: two identical α chains and two identical β chains, all four held together by noncovalent interactions. Each α subunit is paired in an identical way with a β subunit within the structure of this multisubunit protein, so that hemoglobin can be considered either a tetramer of four polypeptide subunits or a dimer of $\alpha\beta$ protomers.

A few proteins contain two or more polypeptide chains linked covalently. For example, the two polypeptide chains of insulin are linked by disulfide bonds. In such cases, the individual polypeptides are not considered subunits but are commonly referred to simply as chains.

We can calculate the approximate number of amino acid residues in a simple protein containing no other chemical constituents by dividing its molecular weight by 110. Although the average molecular weight of the 20 common amino acids is about 138, the smaller amino acids predominate in most proteins. If we take into account the proportions in which the various amino acids occur in proteins (Table 3–1), the average molecular weight of protein amino acids is nearer to 128. Because a molecule of water (M_r 18) is removed to create each peptide bond, the average molecular weight of an amino acid residue in a protein is about $128 - 18 = 110$.

Polypeptides Have Characteristic Amino Acid Compositions

Hydrolysis of peptides or proteins with acid yields a mixture of free α-amino acids. When completely hydrolyzed, each type of protein yields a characteristic proportion or mixture of the different amino acids. The 20 common amino acids almost never occur in equal amounts in a protein. Some amino acids may occur only once or not at all in a given type of protein; others may occur in large numbers. Table 3–3 shows the composition of the amino acid mixtures obtained on complete hydrolysis of bovine cytochrome *c* and chymotrypsinogen, the inactive precursor of the digestive enzyme chymotrypsin. These two proteins, with very different functions, also differ significantly in the relative numbers of each kind of amino acid they contain.

Complete hydrolysis alone is not sufficient for an exact analysis of amino acid composition, however, because some side reactions occur during the procedure. For example, the amide bonds in the side chains of asparagine and glutamine are cleaved by acid treatment, yielding aspartate and glutamate, respectively. The side chain of tryptophan is almost completely degraded by acid hydrolysis, and small amounts of serine, threonine,

TABLE 3–2 Molecular Data on Some Proteins

	Molecular weight	*Number of residues*	*Number of polypeptide chains*
Cytochrome *c* (human)	13,000	104	1
Ribonuclease A (bovine pancreas)	13,700	124	1
Lysozyme (chicken egg white)	13,930	129	1
Myoglobin (equine heart)	16,890	153	1
Chymotrypsin (bovine pancreas)	21,600	241	3
Chymotrypsinogen (bovine)	22,000	245	1
Hemoglobin (human)	64,500	574	4
Serum albumin (human)	68,500	609	1
Hexokinase (yeast)	102,000	972	2
RNA polymerase (*E. coli*)	450,000	4,158	5
Apolipoprotein B (human)	513,000	4,536	1
Glutamine synthetase (*E. coli*)	619,000	5,628	12
Titin (human)	2,993,000	26,926	1

TABLE 3–3 Amino Acid Composition of Two Proteins

Amino acid	Number of residues per molecule of protein*	
	Bovine cytochrome c	Bovine chymotrypsinogen
Ala	6	22
Arg	2	4
Asn	5	15
Asp	3	8
Cys	2	10
Gln	3	10
Glu	9	5
Gly	14	23
His	3	2
Ile	6	10
Leu	6	19
Lys	18	14
Met	2	2
Phe	4	6
Pro	4	9
Ser	1	28
Thr	8	23
Trp	1	8
Tyr	4	4
Val	3	23
Total	104	245

*In some common analyses, such as acid hydrolysis, Asp and Asn are not readily distinguished from each other and are together designated Asx (or B). Similarly, when Glu and Gln cannot be distinguished, they are together designated Glx (or Z). In addition, Trp is destroyed. Additional procedures must be employed to obtain an accurate assessment of complete amino acid content.

and tyrosine are also lost. When a precise amino acid composition is required, biochemists use additional procedures to resolve the ambiguities that remain from acid hydrolysis.

Some Proteins Contain Chemical Groups Other Than Amino Acids

Many proteins, for example the enzymes ribonuclease A and chymotrypsinogen, contain only amino acid residues and no other chemical constituents; these are considered simple proteins. However, some proteins contain permanently associated chemical components in addition to amino acids; these are called **conjugated proteins.** The non–amino acid part of a conjugated protein is usually called its **prosthetic group.** Conjugated proteins are classified on the basis of the chemical nature of their prosthetic groups (Table 3–4); for example, **lipoproteins** contain lipids, **glycoproteins** contain sugar groups, and **metalloproteins** contain a specific metal. A number of proteins contain more than one prosthetic group. Usually the prosthetic group plays an important role in the protein's biological function.

TABLE 3–4 Conjugated Proteins

Class	Prosthetic group	Example
Lipoproteins	Lipids	β_1-Lipoprotein of blood
Glycoproteins	Carbohydrates	Immunoglobulin G
Phosphoproteins	Phosphate groups	Casein of milk
Hemoproteins	Heme (iron porphyrin)	Hemoglobin
Flavoproteins	Flavin nucleotides	Succinate dehydrogenase
Metalloproteins	Iron	Ferritin
	Zinc	Alcohol dehydrogenase
	Calcium	Calmodulin
	Molybdenum	Dinitrogenase
	Copper	Plastocyanin

There Are Several Levels of Protein Structure

For large macromolecules such as proteins, the tasks of describing and understanding structure are approached at several levels of complexity, arranged in a kind of conceptual hierarchy. Four levels of protein structure are commonly defined (Fig. 3–16). A description of all covalent bonds (mainly peptide bonds and disulfide bonds) linking amino acid residues in a polypeptide chain is its **primary structure.** The most important element of primary structure is the *sequence* of amino acid residues. **Secondary structure** refers to particularly stable arrangements of amino acid residues giving rise to recurring structural patterns. **Tertiary structure** describes all aspects of the three-dimensional folding of a polypeptide. When a protein has two or more polypeptide subunits, their arrangement in space is referred to as **quaternary structure.** Primary structure is the focus of Section 3.4; the higher levels of structure are discussed in Chapter 4.

SUMMARY 3.2 Peptides and Proteins

- Amino acids can be joined covalently through peptide bonds to form peptides and proteins. Cells generally contain thousands of different proteins, each with a different biological activity.
- Proteins can be very long polypeptide chains of 100 to several thousand amino acid residues. However, some naturally occurring peptides have only a few amino acid residues. Some proteins are composed of several noncovalently

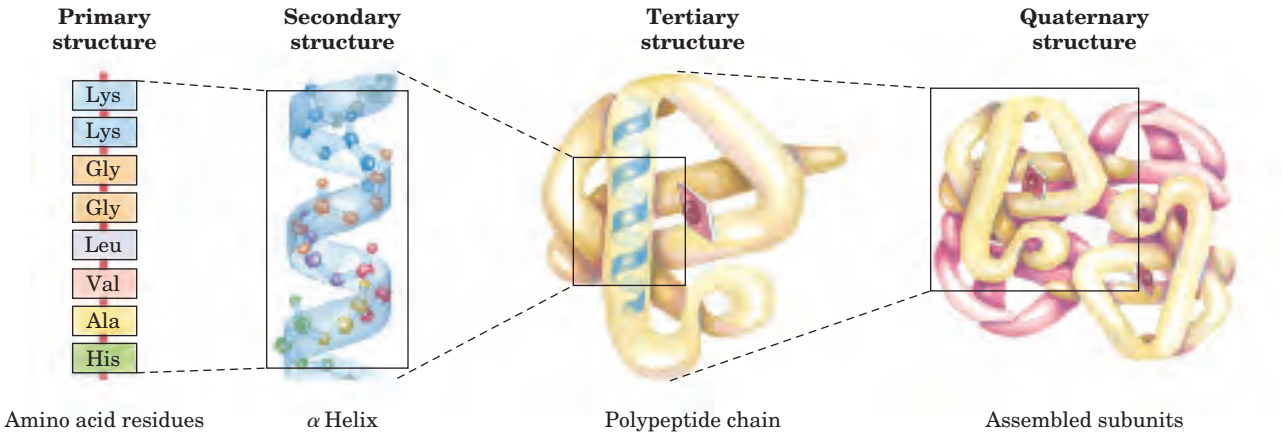


FIGURE 3–16 Levels of structure in proteins. The *primary structure* consists of a sequence of amino acids linked together by peptide bonds and includes any disulfide bonds. The resulting polypeptide can be coiled into units of *secondary structure,* such as an α helix. The helix is a part of the *tertiary structure* of the folded polypeptide, which is itself one of the subunits that make up the *quaternary structure* of the multisubunit protein, in this case hemoglobin.

associated polypeptide chains, called subunits. Simple proteins yield only amino acids on hydrolysis; conjugated proteins contain in addition some other component, such as a metal or organic prosthetic group.

- The sequence of amino acids in a protein is characteristic of that protein and is called its primary structure. This is one of four generally recognized levels of protein structure.

3.3 Working with Proteins

Our understanding of protein structure and function has been derived from the study of many individual proteins. To study a protein in detail, the researcher must be able to separate it from other proteins and must have the techniques to determine its properties. The necessary methods come from protein chemistry, a discipline as old as biochemistry itself and one that retains a central position in biochemical research.

Proteins Can Be Separated and Purified

A pure preparation is essential before a protein's properties and activities can be determined. Given that cells contain thousands of different kinds of proteins, how can one protein be purified? Methods for separating proteins take advantage of properties that vary from one protein to the next, including size, charge, and binding properties.

The source of a protein is generally tissue or microbial cells. The first step in any protein purification procedure is to break open these cells, releasing their proteins into a solution called a **crude extract.** If necessary, differential centrifugation can be used to prepare subcellular fractions or to isolate specific organelles (see Fig. 1–8).

Once the extract or organelle preparation is ready, various methods are available for purifying one or more of the proteins it contains. Commonly, the extract is subjected to treatments that separate the proteins into different fractions based on a property such as size or charge, a process referred to as **fractionation.** Early fractionation steps in a purification utilize differences in protein solubility, which is a complex function of pH, temperature, salt concentration, and other factors. The solubility of proteins is generally lowered at high salt concentrations, an effect called "salting out." The addition of a salt in the right amount can selectively precipitate some proteins, while others remain in solution. Ammonium sulfate ($(NH_4)_2SO_4$) is often used for this purpose because of its high solubility in water.

A solution containing the protein of interest often must be further altered before subsequent purification steps are possible. For example, **dialysis** is a procedure that separates proteins from solvents by taking advantage of the proteins' larger size. The partially purified extract is placed in a bag or tube made of a semipermeable membrane. When this is suspended in a much larger volume of buffered solution of appropriate ionic strength, the membrane allows the exchange of salt and buffer but not proteins. Thus dialysis retains large proteins within the membranous bag or tube while allowing the concentration of other solutes in the protein preparation to change until they come into equilibrium with the solution outside the membrane. Dialysis might be used, for example, to remove ammonium sulfate from the protein preparation.

The most powerful methods for fractionating proteins make use of **column chromatography,** which takes advantage of differences in protein charge, size,

binding affinity, and other properties (Fig. 3–17). A porous solid material with appropriate chemical properties (the stationary phase) is held in a column, and a buffered solution (the mobile phase) percolates through it. The protein-containing solution, layered on the top of the column, percolates through the solid matrix as an ever-expanding band within the larger mobile phase (Fig. 3–17). Individual proteins migrate faster or more slowly through the column depending on their properties. For example, in **cation-exchange chromatography** (Fig. 3–18a), the solid matrix has negatively charged groups. In the mobile phase, proteins with a net positive charge migrate through the matrix more slowly than those with a net negative charge, because the migration of the former is retarded more by interaction with the stationary phase. The two types of protein can separate into two distinct bands. The expansion of the protein band in the mobile phase (the protein solution) is caused both by separation of proteins with different properties and by diffusional spreading. As the length of the column increases, the resolution of two types of protein with different net charges generally improves. However, the rate at which the protein solution can flow through the column usually decreases with column length. And as the length of time spent on the column increases, the resolution can decline as a result of diffusional spreading within each protein band.

Figure 3–18 shows two other variations of column chromatography in addition to ion exchange. **Size-exclusion chromatography** separates proteins according to size. In this method, large proteins emerge from the column sooner than small ones—a somewhat counterintuitive result. The solid phase consists of beads with engineered pores or cavities of a particular size. Large proteins cannot enter the cavities, and so take a short (and rapid) path through the column, around the beads. Small proteins enter the cavities, and migrate through the column more slowly as a result (Fig. 3–18b). **Affinity chromatography** is based on the binding affinity of a protein. The beads in the column have a covalently attached chemical group. A protein with affinity for this particular chemical group will bind to the beads in the column, and its migration will be retarded as a result (Fig. 3–18c).

A modern refinement in chromatographic methods is **HPLC,** or **high-performance liquid chromatography.** HPLC makes use of high-pressure pumps that speed the movement of the protein molecules down the column, as well as higher-quality chromatographic materials that can withstand the crushing force of the pressurized flow. By reducing the transit time on the column, HPLC can limit diffusional spreading of protein bands and thus greatly improve resolution.

The approach to purification of a protein that has not previously been isolated is guided both by established precedents and by common sense. In most cases, several different methods must be used sequentially to purify a protein completely. The choice of method is

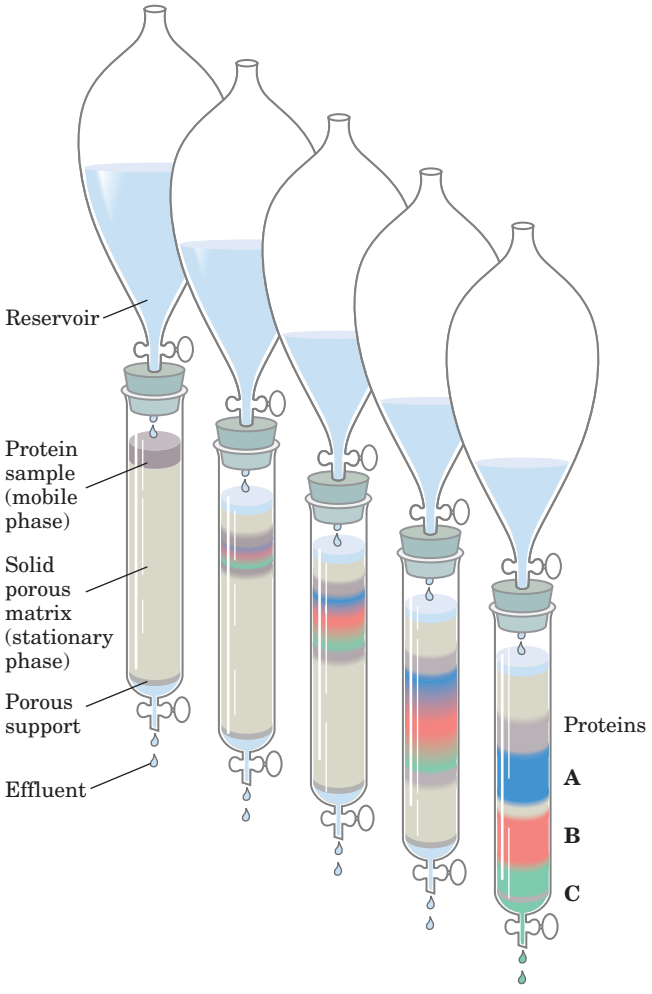


FIGURE 3–17 Column chromatography. The standard elements of a chromatographic column include a solid, porous material supported inside a column, generally made of plastic or glass. The solid material (matrix) makes up the stationary phase through which flows a solution, the mobile phase. The solution that passes out of the column at the bottom (the effluent) is constantly replaced by solution supplied from a reservoir at the top. The protein solution to be separated is layered on top of the column and allowed to percolate into the solid matrix. Additional solution is added on top. The protein solution forms a band within the mobile phase that is initially the depth of the protein solution applied to the column. As proteins migrate through the column, they are retarded to different degrees by their different interactions with the matrix material. The overall protein band thus widens as it moves through the column. Individual types of proteins (such as A, B, and C, shown in blue, red, and green) gradually separate from each other, forming bands within the broader protein band. Separation improves (resolution increases) as the length of the column increases. However, each individual protein band also broadens with time due to diffusional spreading, a process that decreases resolution. In this example, protein A is well separated from B and C, but diffusional spreading prevents complete separation of B and C under these conditions.

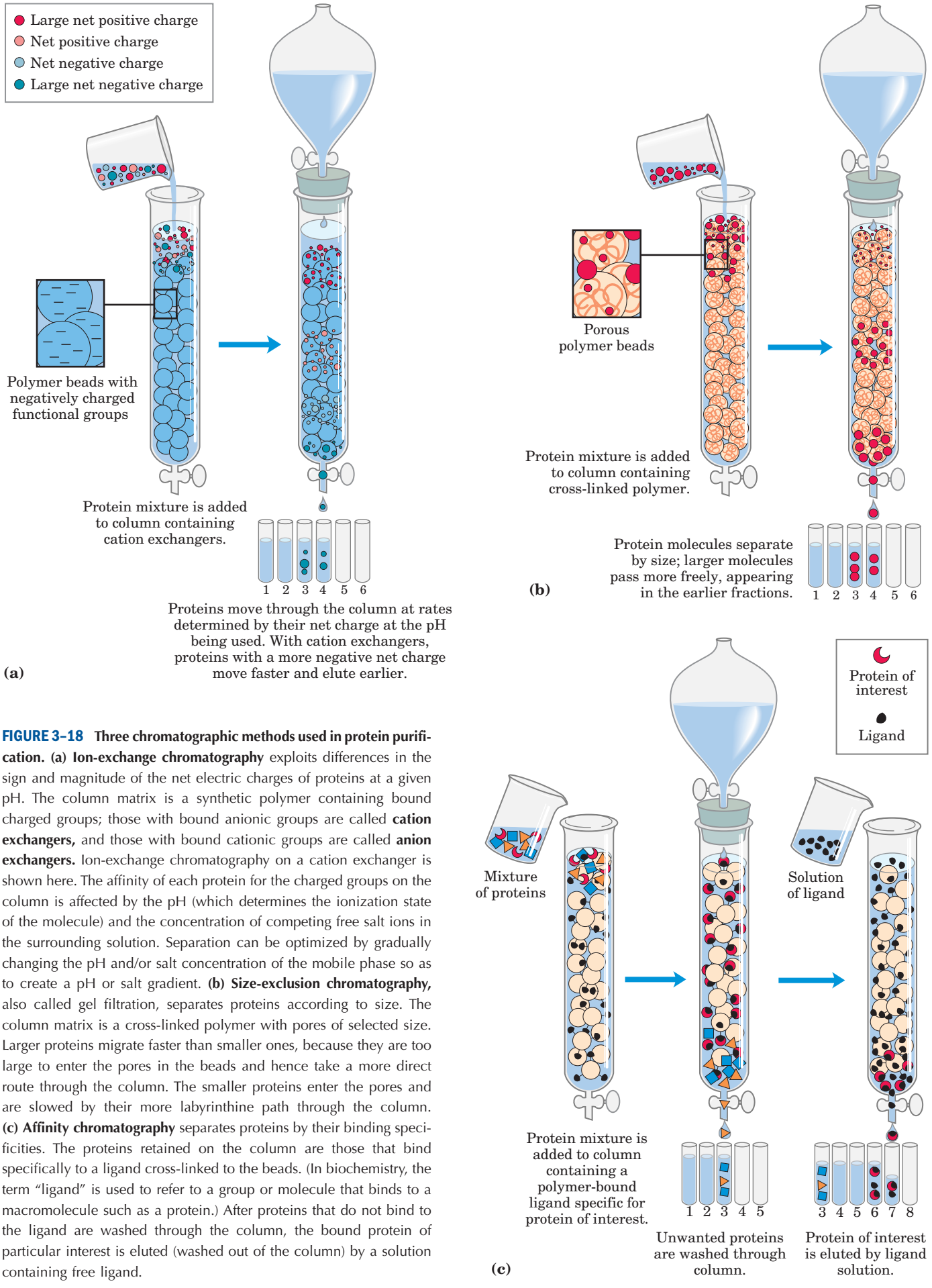


FIGURE 3–18 **Three chromatographic methods used in protein purification. (a) Ion-exchange chromatography** exploits differences in the sign and magnitude of the net electric charges of proteins at a given pH. The column matrix is a synthetic polymer containing bound charged groups; those with bound anionic groups are called **cation exchangers,** and those with bound cationic groups are called **anion exchangers.** Ion-exchange chromatography on a cation exchanger is shown here. The affinity of each protein for the charged groups on the column is affected by the pH (which determines the ionization state of the molecule) and the concentration of competing free salt ions in the surrounding solution. Separation can be optimized by gradually changing the pH and/or salt concentration of the mobile phase so as to create a pH or salt gradient. **(b) Size-exclusion chromatography,** also called gel filtration, separates proteins according to size. The column matrix is a cross-linked polymer with pores of selected size. Larger proteins migrate faster than smaller ones, because they are too large to enter the pores in the beads and hence take a more direct route through the column. The smaller proteins enter the pores and are slowed by their more labyrinthine path through the column. **(c) Affinity chromatography** separates proteins by their binding specificities. The proteins retained on the column are those that bind specifically to a ligand cross-linked to the beads. (In biochemistry, the term "ligand" is used to refer to a group or molecule that binds to a macromolecule such as a protein.) After proteins that do not bind to the ligand are washed through the column, the bound protein of particular interest is eluted (washed out of the column) by a solution containing free ligand.

TABLE 3–5 A Purification Table for a Hypothetical Enzyme

Procedure or step	*Fraction volume (ml)*	*Total protein (mg)*	*Activity (units)*	*Specific activity (units/mg)*
1. Crude cellular extract	1,400	10,000	100,000	10
2. Precipitation with ammonium sulfate	280	3,000	96,000	32
3. Ion-exchange chromatography	90	400	80,000	200
4. Size-exclusion chromatography	80	100	60,000	600
5. Affinity chromatography	6	3	45,000	15,000

Note: All data represent the status of the sample *after* the designated procedure has been carried out. Activity and specific activity are defined on page 94.

somewhat empirical, and many protocols may be tried before the most effective one is found. Trial and error can often be minimized by basing the procedure on purification techniques developed for similar proteins. Published purification protocols are available for many thousands of proteins. Common sense dictates that inexpensive procedures such as salting out be used first, when the total volume and the number of contaminants are greatest. Chromatographic methods are often impractical at early stages, because the amount of chromatographic medium needed increases with sample size. As each purification step is completed, the sample size generally becomes smaller (Table 3–5), making it feasible to use more sophisticated (and expensive) chromatographic procedures at later stages.

Proteins Can Be Separated and Characterized by Electrophoresis

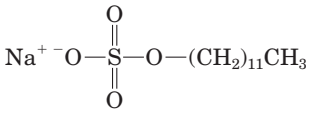
Another important technique for the separation of proteins is based on the migration of charged proteins in an electric field, a process called **electrophoresis.** These procedures are not generally used to purify proteins in large amounts, because simpler alternatives are usually available and electrophoretic methods often adversely affect the structure and thus the function of proteins. Electrophoresis is, however, especially useful as an analytical method. Its advantage is that proteins can be visualized as well as separated, permitting a researcher to estimate quickly the number of different proteins in a mixture or the degree of purity of a particular protein preparation. Also, electrophoresis allows determination of crucial properties of a protein such as its isoelectric point and approximate molecular weight.

Electrophoresis of proteins is generally carried out in gels made up of the cross-linked polymer polyacrylamide (Fig. 3–19). The polyacrylamide gel acts as a molecular sieve, slowing the migration of proteins approximately in proportion to their charge-to-mass ratio. Migration may also be affected by protein shape. In electrophoresis, the force moving the macromolecule is the electrical potential, E. The electrophoretic mobility of the molecule, μ, is the ratio of the velocity of the particle molecule, V, to the electrical potential. Electrophoretic mobility is also equal to the net charge of the molecule, Z, divided by the frictional coefficient, f, which reflects in part a protein's shape. Thus:

$$\mu = \frac{V}{E} = \frac{Z}{f}$$

The migration of a protein in a gel during electrophoresis is therefore a function of its size and its shape.

An electrophoretic method commonly employed for estimation of purity and molecular weight makes use of the detergent **sodium dodecyl sulfate (SDS).**



Sodium dodecyl sulfate (SDS)

SDS binds to most proteins in amounts roughly proportional to the molecular weight of the protein, about one molecule of SDS for every two amino acid residues. The bound SDS contributes a large net negative charge, rendering the intrinsic charge of the protein insignificant and conferring on each protein a similar charge-to-mass ratio. In addition, the native conformation of a protein is altered when SDS is bound, and most proteins assume a similar shape. Electrophoresis in the presence of SDS therefore separates proteins almost exclusively on the basis of mass (molecular weight), with smaller polypeptides migrating more rapidly. After electrophoresis, the proteins are visualized by adding a dye such as Coomassie blue, which binds to proteins but not to the gel itself (Fig. 3–19b). Thus, a researcher can monitor the progress of a protein purification procedure as the number of protein bands visible on the gel decreases after each new fractionation step. When compared with the positions to which proteins of known molecular weight migrate in the gel, the position of an unidentified protein can provide an excellent measure of its molecular weight (Fig. 3–20). If the protein has two or more different subunits, the subunits will generally be separated by the SDS treatment and a separate band will appear for each. SDS Gel Electrophoresis

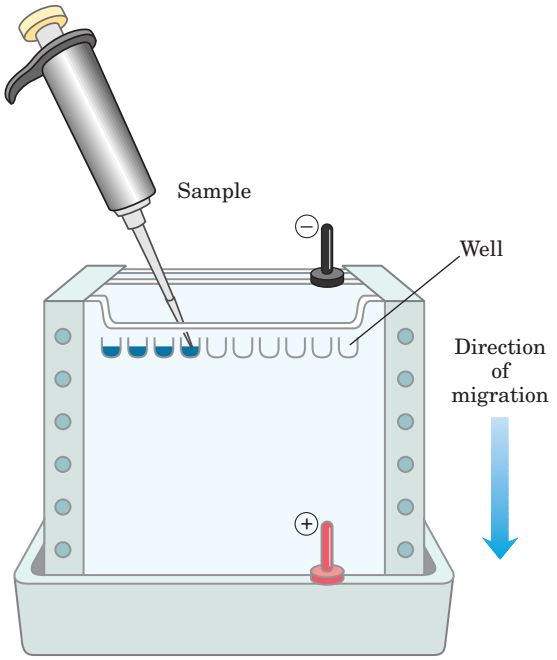
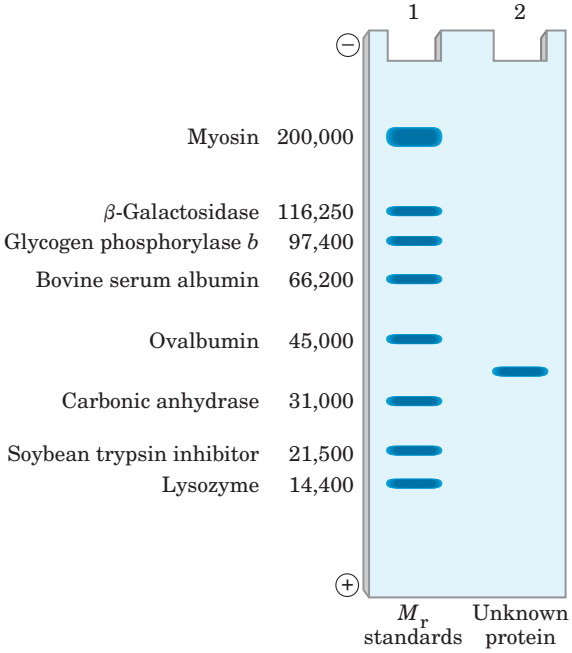


FIGURE 3–19 Electrophoresis. (a) Different samples are loaded in wells or depressions at the top of the polyacrylamide gel. The proteins move into the gel when an electric field is applied. The gel minimizes convection currents caused by small temperature gradients, as well as protein movements other than those induced by the electric field. **(b)** Proteins can be visualized after electrophoresis by treating the gel with a stain such as Coomassie blue, which binds to the proteins but not to the gel itself. Each band on the gel represents a different protein (or protein subunit); smaller proteins move through the gel more rapidly than larger proteins and therefore are found nearer the bottom of the gel. This gel illustrates the purification of the enzyme RNA polymerase from *E. coli*. The first lane shows the proteins present in the crude cellular extract. Successive lanes (left to right) show the proteins present after each purification step. The purified protein contains four subunits, as seen in the last lane on the right.

Isoelectric focusing is a procedure used to determine the isoelectric point (pI) of a protein (Fig. 3–21). A pH gradient is established by allowing a mixture of low molecular weight organic acids and bases (ampholytes; p. 81) to distribute themselves in an electric field generated across the gel. When a protein mixture is applied, each protein migrates until it reaches the pH that matches its pI (Table 3–6). Proteins with different isoelectric points are thus distributed differently throughout the gel.

Combining isoelectric focusing and SDS electrophoresis sequentially in a process called **two-dimensional**



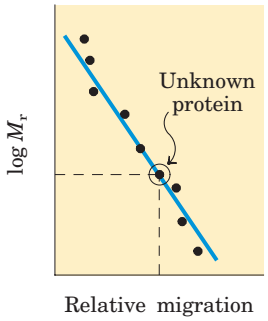


FIGURE 3–20 Estimating the molecular weight of a protein. The electrophoretic mobility of a protein on an SDS polyacrylamide gel is related to its molecular weight, M_r. **(a)** Standard proteins of known molecular weight are subjected to electrophoresis (lane 1). These marker proteins can be used to estimate the molecular weight of an unknown protein (lane 2). **(b)** A plot of log M_r of the marker proteins versus relative migration during electrophoresis is linear, which allows the molecular weight of the unknown protein to be read from the graph.

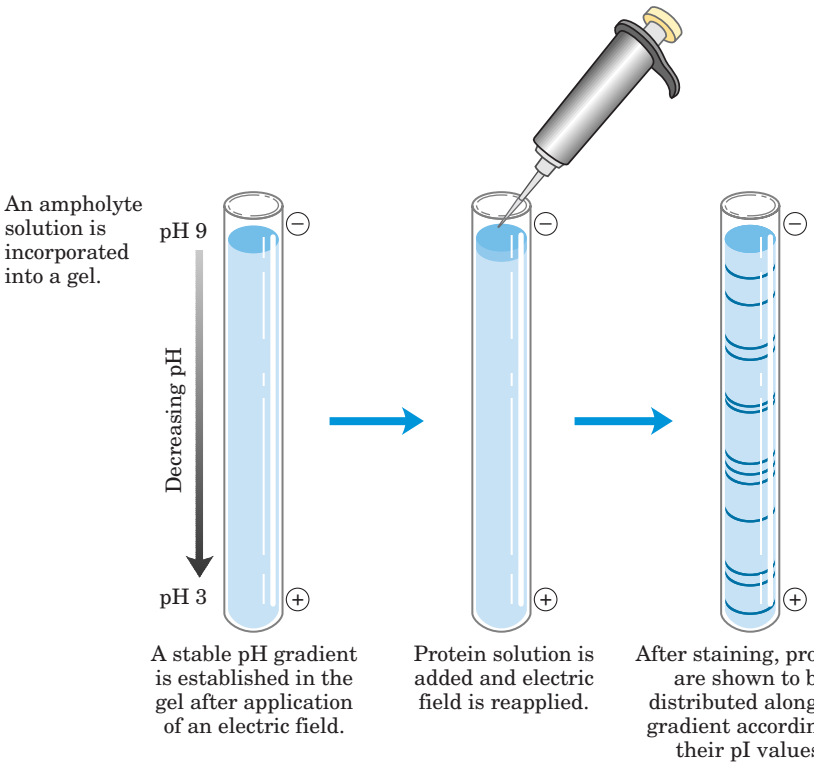


FIGURE 3–21 Isoelectric focusing. This technique separates proteins according to their isoelectric points. A stable pH gradient is established in the gel by the addition of appropriate ampholytes. A protein mixture is placed in a well on the gel. With an applied electric field, proteins enter the gel and migrate until each reaches a pH equivalent to its pI. Remember that when pH = pI, the net charge of a protein is zero.

electrophoresis permits the resolution of complex mixtures of proteins (Fig. 3–22). This is a more sensitive analytical method than either electrophoretic method alone. Two-dimensional electrophoresis separates proteins of identical molecular weight that differ in pI, or proteins with similar pI values but different molecular weights.

Unseparated Proteins Can Be Quantified

To purify a protein, it is essential to have a way of detecting and quantifying that protein in the presence of many other proteins at each stage of the procedure. Often, purification must proceed in the absence of any information about the size and physical properties of the protein or about the fraction of the total protein mass it represents in the extract. For proteins that are enzymes, the amount in a given solution or tissue extract can be measured, or assayed, in terms of the catalytic effect the enzyme produces—that is, the *increase* in the rate at which its substrate is converted to reaction products when the enzyme is present. For this purpose one must know (1) the overall equation of the reaction catalyzed, (2) an analytical procedure for determining the disappearance of the substrate or the appearance of a reaction product, (3) whether the enzyme requires cofactors such as metal ions or coenzymes, (4) the dependence of the enzyme activity on substrate concentration, (5) the optimum pH, and (6) a temperature zone in which the enzyme is stable and has high activity. Enzymes are usually assayed at their optimum pH and at some convenient temperature within the range 25 to 38 °C. Also, very high substrate concentrations are generally used so that the initial reaction rate, measured experimentally, is proportional to enzyme concentration (Chapter 6).

By international agreement, 1.0 unit of enzyme activity is defined as the amount of enzyme causing transformation of 1.0 μmol of substrate per minute at 25 °C under optimal conditions of measurement. The term **activity** refers to the total units of enzyme in a solution. The **specific activity** is the number of enzyme units per milligram of total protein (Fig. 3–23). The specific activity is a measure of enzyme purity: it increases during purification of an enzyme and becomes maximal and constant when the enzyme is pure (Table 3–5).

TABLE 3–6 The Isoelectric Points of Some Proteins

Protein	*pI*
Pepsin	<1.0
Egg albumin	4.6
Serum albumin	4.9
Urease	5.0
β-Lactoglobulin	5.2
Hemoglobin	6.8
Myoglobin	7.0
Chymotrypsinogen	9.5
Cytochrome *c*	10.7
Lysozyme	11.0

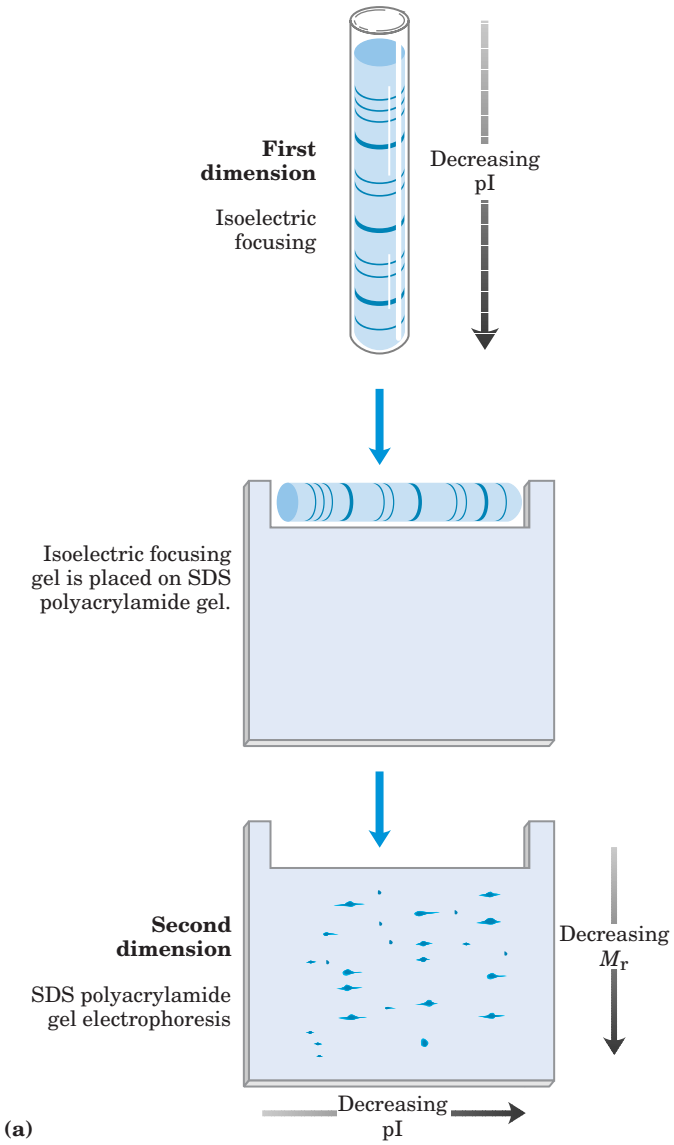




FIGURE 3–22 Two-dimensional electrophoresis. (a) Proteins are first separated by isoelectric focusing in a cylindrical gel. The gel is then laid horizontally on a second, slab-shaped gel, and the proteins are separated by SDS polyacrylamide gel electrophoresis. Horizontal separation reflects differences in pI; vertical separation reflects differences in molecular weight. **(b)** More than 1,000 different proteins from *E. coli* can be resolved using this technique.

After each purification step, the activity of the preparation (in units of enzyme activity) is assayed, the total amount of protein is determined independently, and the ratio of the two gives the specific activity. Activity and total protein generally decrease with each step. Activity decreases because some loss always occurs due to inactivation or nonideal interactions with chromatographic materials or other molecules in the solution. Total protein decreases because the objective is to remove as much unwanted or nonspecific protein as possible. In a successful step, the loss of nonspecific protein is much greater than the loss of activity; therefore, specific activity increases even as total activity falls. The data are then assembled in a purification table similar to Table 3–5. A protein is generally considered pure when further purification steps fail to increase specific activity and when only a single protein species can be detected (for example, by electrophoresis).

For proteins that are not enzymes, other quantification methods are required. Transport proteins can be assayed by their binding to the molecule they transport, and hormones and toxins by the biological effect they produce; for example, growth hormones will stimulate the growth of certain cultured cells. Some structural proteins represent such a large fraction of a tissue mass that they can be readily extracted and purified without a functional assay. The approaches are as varied as the proteins themselves.

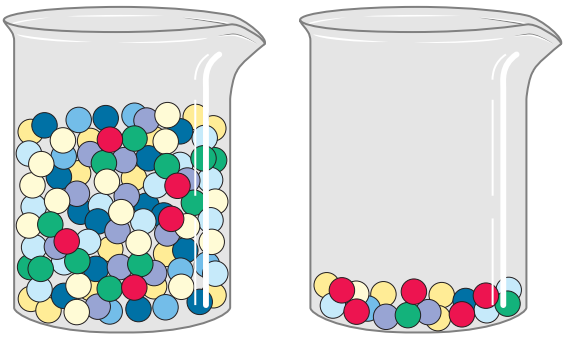


FIGURE 3–23 Activity versus specific activity. The difference between these two terms can be illustrated by considering two beakers of marbles. The beakers contain the same number of red marbles, but different numbers of marbles of other colors. If the marbles represent proteins, both beakers contain the same *activity* of the protein represented by the red marbles. The second beaker, however, has the higher *specific activity* because here the red marbles represent a much higher fraction of the total.

SUMMARY 3.3 Working with Proteins

- Proteins are separated and purified by taking advantage of differences in their properties. Proteins can be selectively precipitated by the addition of certain salts. A wide range of chromatographic procedures makes use of differences in size, binding affinities, charge, and other properties. These include ion-exchange, size-exclusion, affinity, and high-performance liquid chromatography.
- Electrophoresis separates proteins on the basis of mass or charge. SDS gel electrophoresis and isoelectric focusing can be used separately or in combination for higher resolution.
- All purification procedures require a method for quantifying or assaying the protein of interest in the presence of other proteins. Purification can be monitored by assaying specific activity.

3.4 The Covalent Structure of Proteins

Purification of a protein is usually only a prelude to a detailed biochemical dissection of its structure and function. What is it that makes one protein an enzyme, another a hormone, another a structural protein, and still another an antibody? How do they differ chemically? The most obvious distinctions are structural, and these distinctions can be approached at every level of structure defined in Figure 3–16.

The differences in primary structure can be especially informative. Each protein has a distinctive number and sequence of amino acid residues. As we shall see in Chapter 4, the primary structure of a protein determines how it folds up into a unique three-dimensional structure, and this in turn determines the function of the protein. Primary structure is the focus of the remainder of this chapter. We first consider empirical clues that amino acid sequence and protein function are closely linked, then describe how amino acid sequence is determined; finally, we outline the many uses to which this information can be put.

The Function of a Protein Depends on Its Amino Acid Sequence

The bacterium *Escherichia coli* produces more than 3,000 different proteins; a human produces 25,000 to 35,000. In both cases, each type of protein has a unique three-dimensional structure and this structure confers a unique function. Each type of protein also has a unique amino acid sequence. Intuition suggests that the amino acid sequence must play a fundamental role in determining the three-dimensional structure of the protein, and ultimately its function, but is this supposition correct? A quick survey of proteins and how they vary in amino acid sequence provides a number of empirical clues that help substantiate the important relationship between amino acid sequence and biological function.

First, as we have already noted, proteins with different functions always have different amino acid sequences. Second, thousands of human genetic diseases have been traced to the production of defective proteins. Perhaps one-third of these proteins are defective because of a single change in their amino acid sequence; hence, if the primary structure is altered, the function of the protein may also be changed. Finally, on comparing functionally similar proteins from different species, we find that these proteins often have similar amino acid sequences. An extreme case is ubiquitin, a 76-residue protein involved in regulating the degradation of other proteins. The amino acid sequence of ubiquitin is identical in species as disparate as fruit flies and humans.

Is the amino acid sequence absolutely fixed, or invariant, for a particular protein? No; some flexibility is possible. An estimated 20% to 30% of the proteins in humans are **polymorphic,** having amino acid sequence variants in the human population. Many of these variations in sequence have little or no effect on the function of the protein. Furthermore, proteins that carry out a broadly similar function in distantly related species can differ greatly in overall size and amino acid sequence.

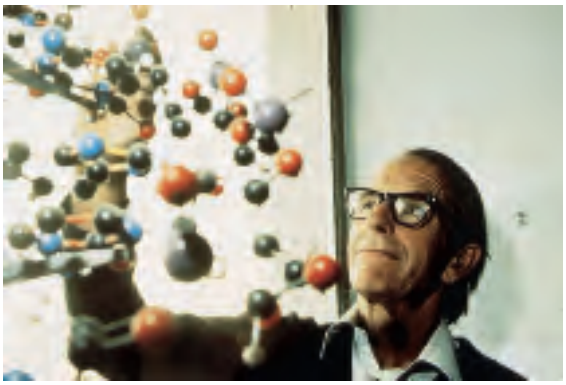
Although the amino acid sequence in some regions of the primary structure might vary considerably without affecting biological function, most proteins contain crucial regions that are essential to their function and whose sequence is therefore conserved. The fraction of the overall sequence that is critical varies from protein to protein, complicating the task of relating sequence to three-dimensional structure, and structure to function. Before we can consider this problem further, however, we must examine how sequence information is obtained.

The Amino Acid Sequences of Millions of Proteins Have Been Determined

Two major discoveries in 1953 were of crucial importance in the history of biochemistry. In that year James D. Watson and Francis Crick deduced the double-helical structure of DNA and proposed a structural basis for its precise replication (Chapter 8). Their proposal illuminated the molecular reality behind the idea of a gene. In that same year, Frederick Sanger worked out the sequence of amino acid residues in the polypeptide chains of the hormone insulin (Fig. 3–24), surprising many researchers who had long thought that elucidation of the amino acid sequence of a polypeptide would be a hopelessly difficult task. It quickly became evident that the nucleotide sequence in DNA and the amino acid sequence in proteins were somehow related. Barely a decade after these discoveries, the role of the nucleotide

sequence of DNA in determining the amino acid sequence of protein molecules was revealed (Chapter 27). An enormous number of protein sequences can now be derived indirectly from the DNA sequences in the rapidly growing genome databases. However, many are still deduced by traditional methods of polypeptide sequencing.

The amino acid sequences of thousands of different proteins from many species have been determined using principles first developed by Sanger. These methods are still in use, although with many variations and improvements in detail. Chemical protein sequencing now complements a growing list of newer methods, providing multiple avenues to obtain amino acid sequence data. Such data are now critical to every area of biochemical investigation.



Frederick Sanger

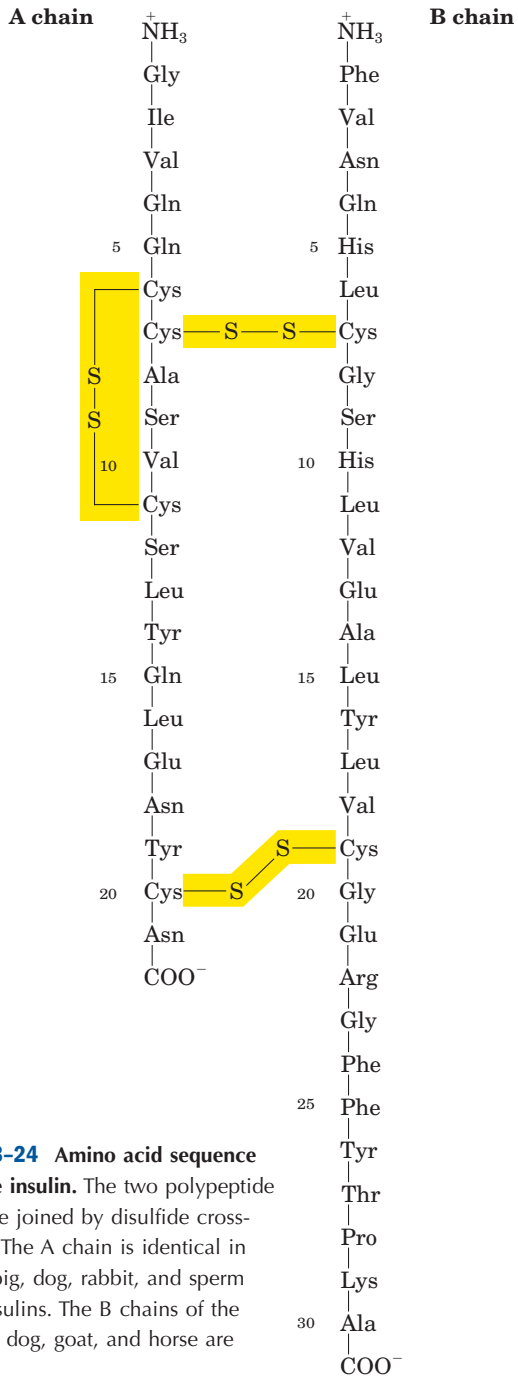
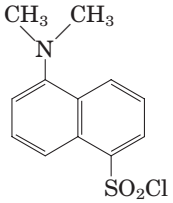


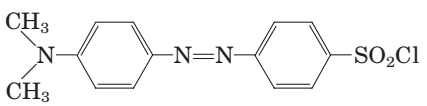
FIGURE 3–24 Amino acid sequence of bovine insulin. The two polypeptide chains are joined by disulfide cross-linkages. The A chain is identical in human, pig, dog, rabbit, and sperm whale insulins. The B chains of the cow, pig, dog, goat, and horse are identical.

Short Polypeptides Are Sequenced Using Automated Procedures

Various procedures are used to analyze protein primary structure. Several protocols are available to label and identify the amino-terminal amino acid residue (Fig. 3–25a). Sanger developed the reagent 1-fluoro-2,4-dinitrobenzene (FDNB) for this purpose; other reagents used to label the amino-terminal residue, dansyl chloride and dabsyl chloride, yield derivatives that are more easily detectable than the dinitrophenyl derivatives. After the amino-terminal residue is labeled with one of these reagents, the polypeptide is hydrolyzed to its constituent amino acids and the labeled amino acid is identified. Because the hydrolysis stage destroys the polypeptide, this procedure cannot be used to sequence a polypeptide beyond its amino-terminal residue. However, it can help determine the number of chemically distinct polypeptides in a protein, provided each has a different amino-terminal residue. For example, two residues—Phe and Gly—would be labeled if insulin (Fig. 3–24) were subjected to this procedure.



Dansyl chloride



Dabsyl chloride

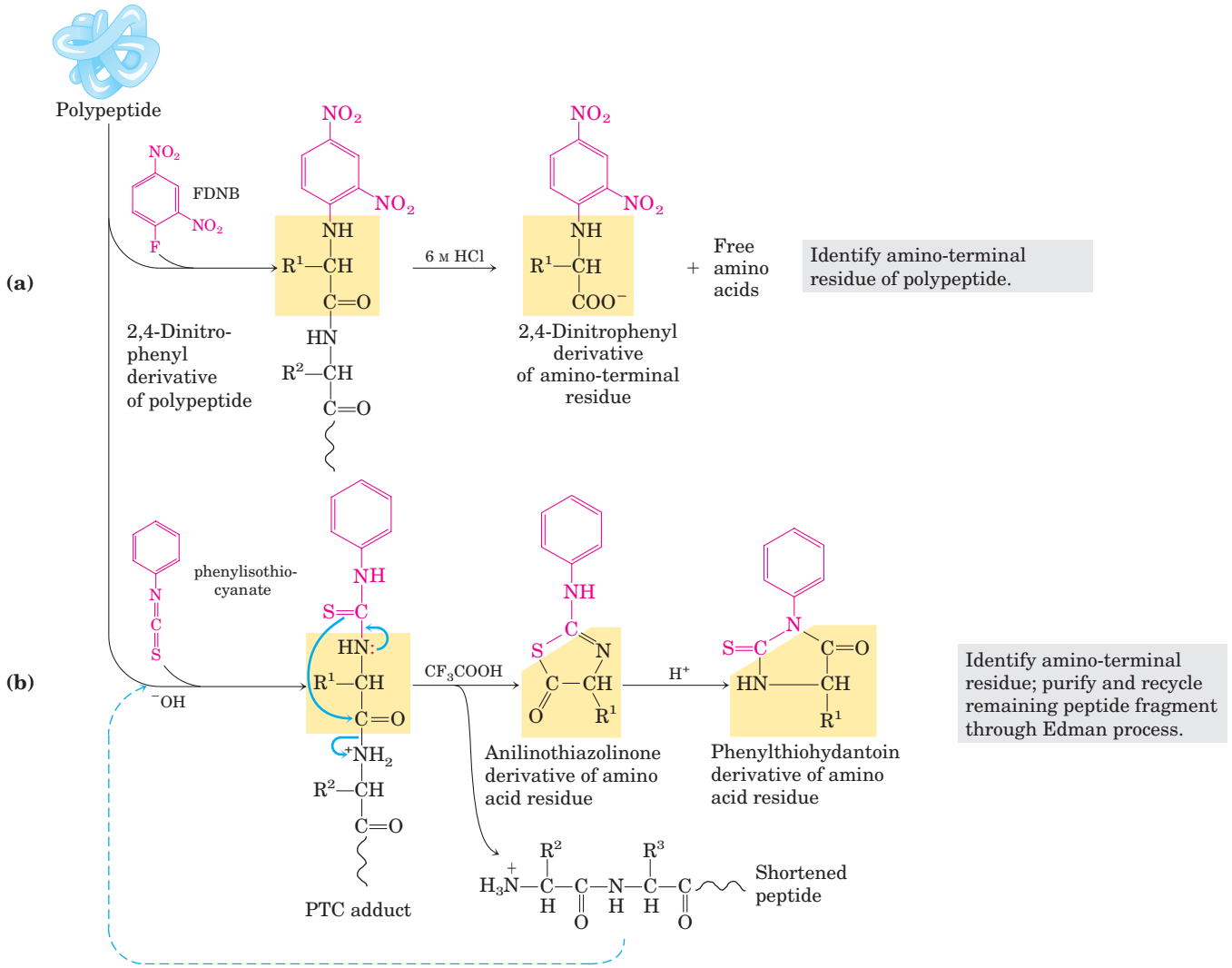


FIGURE 3–25 **Steps in sequencing a polypeptide.** **(a)** Identification of the amino-terminal residue can be the first step in sequencing a polypeptide. Sanger's method for identifying the amino-terminal residue is shown here. **(b)** The Edman degradation procedure reveals the entire sequence of a peptide. For shorter peptides, this method alone readily yields the entire sequence, and step **(a)** is often omitted. Step **(a)** is useful in the case of larger polypeptides, which are often fragmented into smaller peptides for sequencing (see Fig. 3–27).

To sequence an entire polypeptide, a chemical method devised by Pehr Edman is usually employed. The **Edman degradation** procedure labels and removes only the amino-terminal residue from a peptide, leaving all other peptide bonds intact (Fig. 3–25b). The peptide is reacted with phenylisothiocyanate under mildly alkaline conditions, which converts the amino-terminal amino acid to a phenylthiocarbamoyl (PTC) adduct. The peptide bond next to the PTC adduct is then cleaved in a step carried out in anhydrous trifluoroacetic acid, with removal of the amino-terminal amino acid as an anilinothiazolinone derivative. The derivatized amino acid is extracted with organic solvents, converted to the more stable phenylthiohydantoin derivative by treatment with aqueous acid, and then identified. The use of sequential reactions carried out under first basic and then acidic conditions provides control over the entire process. Each reaction with the amino-terminal amino acid can go essentially to completion without affecting any of the other peptide bonds in the peptide. After removal and identification of the amino-terminal residue, the *new* amino-terminal residue so exposed can be labeled, removed, and identified through the same series of reactions. This procedure is repeated until the entire sequence is determined. The Edman degradation is carried out on a machine, called a **sequenator,** that mixes reagents in the proper proportions, separates the products, identifies them, and records the results. These methods are extremely sensitive. Often, the complete amino acid sequence can be determined starting with only a few micrograms of protein.

The length of polypeptide that can be accurately sequenced by the Edman degradation depends on the

efficiency of the individual chemical steps. Consider a peptide beginning with the sequence Gly–Pro–Lys– at its amino terminus. If glycine were removed with 97% efficiency, 3% of the polypeptide molecules in the solution would retain a Gly residue at their amino terminus. In the second Edman cycle, 97% of the liberated amino acids would be proline, and 3% glycine, while 3% of the polypeptide molecules would retain Gly (0.1%) or Pro (2.9%) residues at their amino terminus. At each cycle, peptides that did not react in earlier cycles would contribute amino acids to an ever-increasing background, eventually making it impossible to determine which amino acid is next in the original peptide sequence. Modern sequenators achieve efficiencies of better than 99% per cycle, permitting the sequencing of more than 50 contiguous amino acid residues in a polypeptide. The primary structure of insulin, worked out by Sanger and colleagues over a period of 10 years, could now be completely determined in a day or two.

Large Proteins Must Be Sequenced in Smaller Segments

The overall accuracy of amino acid sequencing generally declines as the length of the polypeptide increases. The very large polypeptides found in proteins must be broken down into smaller pieces to be sequenced efficiently. There are several steps in this process. First, the protein is cleaved into a set of specific fragments by chemical or enzymatic methods. If any disulfide bonds are present, they must be broken. Each fragment is purified, then sequenced by the Edman procedure. Finally, the order in which the fragments appear in the original protein is determined and disulfide bonds (if any) are located.

Breaking Disulfide Bonds Disulfide bonds interfere with the sequencing procedure. A cystine residue (Fig. 3–7) that has one of its peptide bonds cleaved by the Edman procedure may remain attached to another polypeptide strand via its disulfide bond. Disulfide bonds also interfere with the enzymatic or chemical cleavage of the polypeptide. Two approaches to irreversible breakage of disulfide bonds are outlined in Figure 3–26.

Cleaving the Polypeptide Chain Several methods can be used for fragmenting the polypeptide chain. Enzymes called **proteases** catalyze the hydrolytic cleavage of peptide bonds. Some proteases cleave only the peptide bond adjacent to particular amino acid residues (Table 3–7) and thus fragment a polypeptide chain in a predictable and reproducible way. A number of chemical reagents also cleave the peptide bond adjacent to specific residues.

Among proteases, the digestive enzyme trypsin catalyzes the hydrolysis of only those peptide bonds in which the carbonyl group is contributed by either a Lys or an Arg residue, regardless of the length or amino acid sequence of the chain. The number of smaller peptides produced by trypsin cleavage can thus be predicted

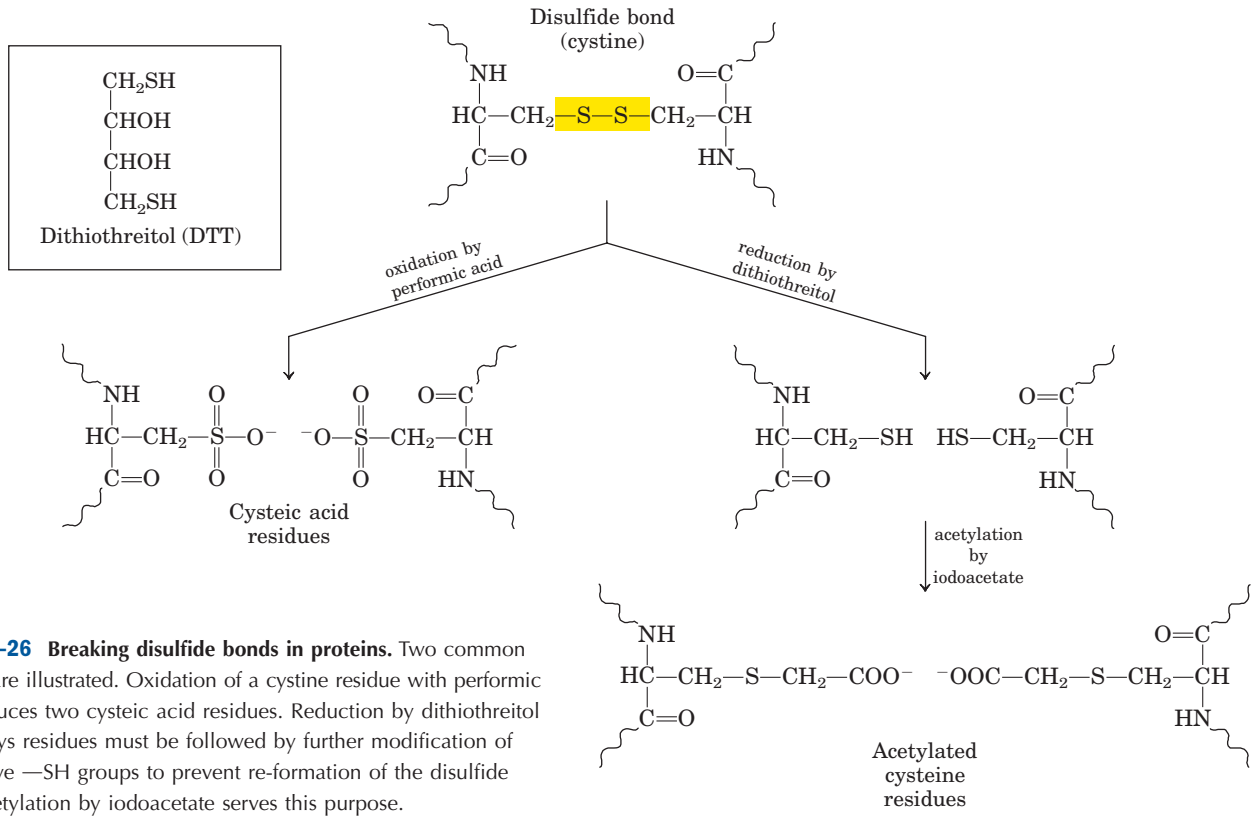


FIGURE 3–26 **Breaking disulfide bonds in proteins.** Two common methods are illustrated. Oxidation of a cystine residue with performic acid produces two cysteic acid residues. Reduction by dithiothreitol to form Cys residues must be followed by further modification of the reactive —SH groups to prevent re-formation of the disulfide bond. Acetylation by iodoacetate serves this purpose.

TABLE 3–7 The Specificity of Some Common Methods for Fragmenting Polypeptide Chains

*Reagent (biological source)**	*Cleavage points*†
Trypsin (bovine pancreas)	Lys, Arg (C)
Submaxillarus protease (mouse submaxillary gland)	Arg (C)
Chymotrypsin (bovine pancreas)	Phe, Trp, Tyr (C)
Staphylococcus aureus V8 protease (bacterium *S. aureus*)	Asp, Glu (C)
Asp-*N*-protease (bacterium *Pseudomonas fragi*)	Asp, Glu (N)
Pepsin (porcine stomach)	Phe, Trp, Tyr (N)
Endoproteinase Lys C (bacterium *Lysobacter enzymogenes*)	Lys (C)
Cyanogen bromide	Met (C)

*All reagents except cyanogen bromide are proteases. All are available from commercial sources.

†Residues furnishing the primary recognition point for the protease or reagent; peptide bond cleavage occurs on either the carbonyl (C) or the amino (N) side of the indicated amino acid residues.

from the total number of Lys or Arg residues in the original polypeptide, as determined by hydrolysis of an intact sample (Fig. 3–27). A polypeptide with five Lys and/or Arg residues will usually yield six smaller peptides on cleavage with trypsin. Moreover, all except one of these will have a carboxyl-terminal Lys or Arg. The fragments produced by trypsin (or other enzyme or chemical) action are then separated by chromatographic or electrophoretic methods.

Sequencing of Peptides Each peptide fragment resulting from the action of trypsin is sequenced separately by the Edman procedure.

Ordering Peptide Fragments The order of the "trypsin fragments" in the original polypeptide chain must now be determined. Another sample of the intact polypeptide is cleaved into fragments using a different enzyme or reagent, one that cleaves peptide bonds at points other than those cleaved by trypsin. For example, cyanogen bromide cleaves only those peptide bonds in which the carbonyl group is contributed by Met. The fragments resulting from this second procedure are then separated and sequenced as before.

The amino acid sequences of each fragment obtained by the two cleavage procedures are examined, with the objective of finding peptides from the second procedure whose sequences establish continuity, because of overlaps, between the fragments obtained by the first cleavage procedure (Fig. 3–27). Overlapping peptides obtained from the second fragmentation yield the correct order of the peptide fragments produced in the first. If the amino-terminal amino acid has been identified before the original cleavage of the protein, this information can be used to establish which fragment is derived from the amino terminus. The two sets of fragments can be compared for possible errors in determining the amino acid sequence of each fragment. If the second cleavage procedure fails to establish continuity between all peptides from the first cleavage, a third or even a fourth cleavage method must be used to obtain a set of peptides that can provide the necessary overlap(s).

Locating Disulfide Bonds If the primary structure includes disulfide bonds, their locations are determined in an additional step after sequencing is completed. A sample of the protein is again cleaved with a reagent such as trypsin, this time without first breaking the disulfide bonds. The resulting peptides are separated by electrophoresis and compared with the original set of peptides generated by trypsin. For each disulfide bond, two of the original peptides will be missing and a new, larger peptide will appear. The two missing peptides represent the regions of the intact polypeptide that are linked by the disulfide bond.

Amino Acid Sequences Can Also Be Deduced by Other Methods

The approach outlined above is not the only way to determine amino acid sequences. New methods based on mass spectrometry permit the sequencing of short polypeptides (20 to 30 amino acid residues) in just a few minutes (Box 3–2). In addition, with the development of rapid DNA sequencing methods (Chapter 8), the elucidation of the genetic code (Chapter 27), and the development of techniques for isolating genes (Chapter 9), researchers can deduce the sequence of a polypeptide by determining the sequence of nucleotides in the gene that codes for it (Fig. 3–28). The techniques used to determine protein and DNA sequences are complementary. When the gene is available, sequencing the DNA can be faster and more accurate than sequencing the protein. Most proteins are now sequenced in this indirect way. If the gene has not been isolated, direct sequencing of peptides is necessary, and this can provide information (the location of disulfide bonds, for example) not available in a DNA sequence. In addition, a knowledge of the amino acid sequence of even a part of a polypeptide can greatly facilitate the isolation of the corresponding gene (Chapter 9).

The array of methods now available to analyze both proteins and nucleic acids is ushering in a new disci-

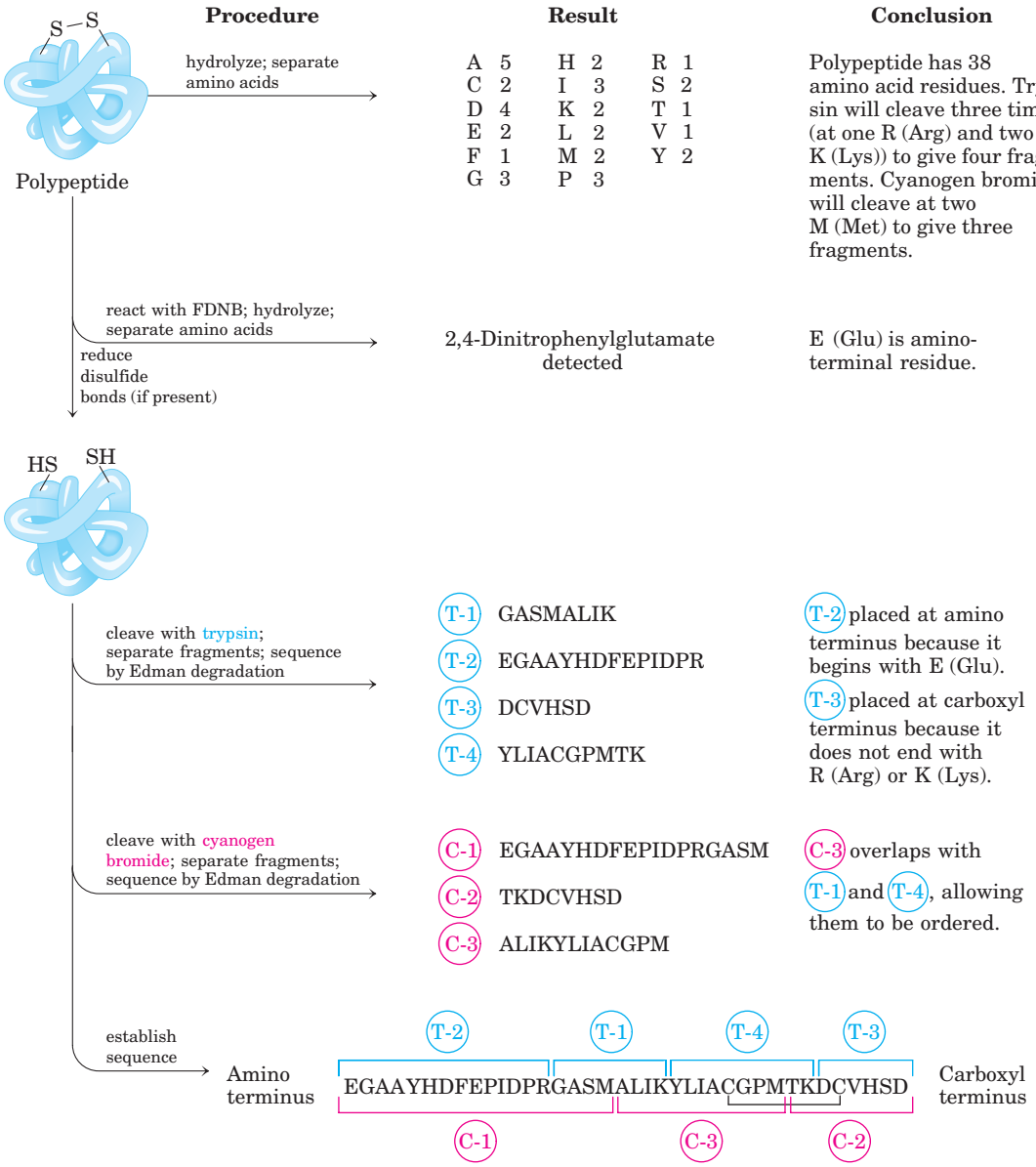


FIGURE 3–27 Cleaving proteins and sequencing and ordering the peptide fragments. First, the amino acid composition and amino-terminal residue of an intact sample are determined. Then any disulfide bonds are broken before fragmenting so that sequencing can proceed efficiently. In this example, there are only two Cys (C) residues and thus only one possibility for location of the disulfide bond. In polypeptides with three or more Cys residues, the position of disulfide bonds can be determined as described in the text. (The one-letter symbols for amino acids are given in Table 3–1.)

pline of "whole cell biochemistry." The complete sequence of an organism's DNA, its genome, is now available for organisms ranging from viruses to bacteria to multicellular eukaryotes (see Table 1–4). Genes are being discovered by the millions, including many that encode proteins with no known function. To describe the entire protein complement encoded by an organism's DNA, researchers have coined the term **proteome.** As described in Chapter 9, the new disciplines of genomics and proteomics are complementing work carried out on cellular intermediary metabolism and nucleic acid metabolism to provide a new and increasingly complete picture of biochemistry at the level of cells and even organisms.

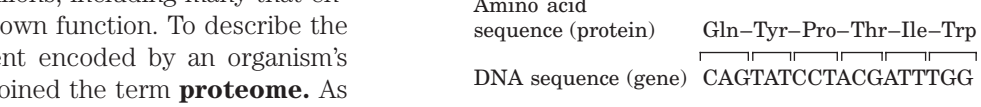


FIGURE 3–28 Correspondence of DNA and amino acid sequences. Each amino acid is encoded by a specific sequence of three nucleotides in DNA. The genetic code is described in detail in Chapter 27.

BOX 3–2 WORKING IN BIOCHEMISTRY

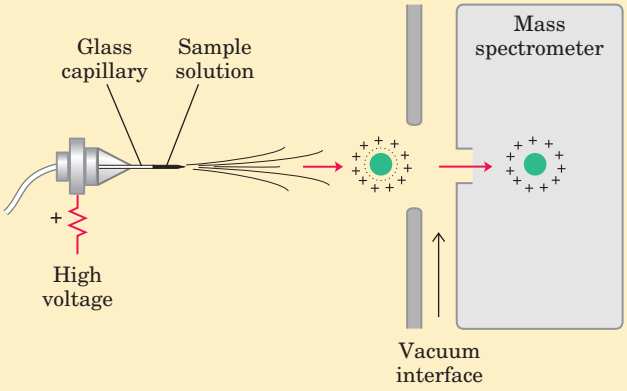
Investigating Proteins with Mass Spectrometry

The mass spectrometer has long been an indispensable tool in chemistry. Molecules to be analyzed, referred to as **analytes,** are first ionized in a vacuum. When the newly charged molecules are introduced into an electric and/or magnetic field, their paths through the field are a function of their mass-to-charge ratio, m/z. This measured property of the ionized species can be used to deduce the mass (M) of the analyte with very high precision.

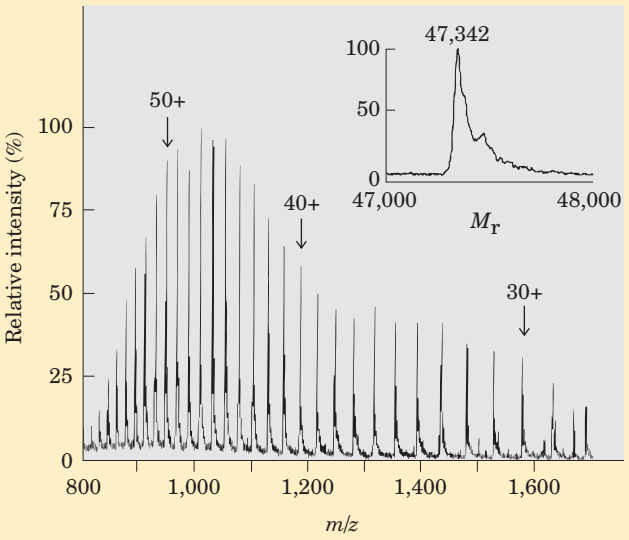
Although mass spectrometry has been in use for many years, it could not be applied to macromolecules such as proteins and nucleic acids. The m/z measurements are made on molecules in the gas phase, and the heating or other treatment needed to transfer a macromolecule to the gas phase usually caused its rapid decomposition. In 1988, two different techniques were developed to overcome this problem. In one, proteins are placed in a light-absorbing matrix. With a short pulse of laser light, the proteins are ionized and then desorbed from the matrix into the vacuum system. This process, known as **matrix-assisted laser desorption/ionization mass spectrometry,** or **MALDI MS,** has been successfully used to measure the mass of a wide range of macromolecules. In a second and equally successful method, macromolecules in solution are forced directly from the liquid to gas phase. A solution of analytes is passed through a charged needle that is kept at a high electrical potential, dispersing the solution into a fine mist of charged microdroplets. The solvent surrounding the macromolecules rapidly evaporates, and the resulting multiply charged macromolecular ions are thus introduced nondestructively into the gas phase. This technique is called **electrospray ionization mass spectrometry,** or **ESI MS.** Protons added during passage through the needle give additional charge to the macromolecule. The m/z of the molecule can be analyzed in the vacuum chamber.

Mass spectrometry provides a wealth of information for proteomics research, enzymology, and protein chemistry in general. The techniques require only miniscule amounts of sample, so they can be readily applied to the small amounts of protein that can be extracted from a two-dimensional electrophoretic gel. The accurately measured molecular mass of a protein is one of the critical parameters in its identification. Once the mass of a protein is accurately known, mass spectrometry is a convenient and accurate method for detecting changes in mass due to the presence of bound cofactors, bound metal ions, covalent modifications, and so on.

The process for determining the molecular mass of a protein with ESI MS is illustrated in Figure 1. As it is injected into the gas phase, a protein acquires a variable number of protons, and thus positive charges, from the solvent. This creates a spectrum of species with different mass-to-charge ratios. Each successive peak corresponds to a species that differs from that



(a)



(b)

FIGURE 1 Electrospray mass spectrometry of a protein. (a) A protein solution is dispersed into highly charged droplets by passage through a needle under the influence of a high-voltage electric field. The droplets evaporate, and the ions (with added protons in this case) enter the mass spectrometer for m/z measurement. The spectrum generated **(b)** is a family of peaks, with each successive peak (from right to left) corresponding to a charged species increased by 1 in both mass and charge. A computer-generated transformation of this spectrum is shown in the inset.

of its neighboring peak by a charge difference of 1 and a mass difference of 1 (1 proton). The mass of the protein can be determined from any two neighboring peaks. The measured m/z of one peak is

$$(m/z)_2 = \frac{M + n_{2X}}{n_2}$$

where M is the mass of the protein, n_2 is the number of charges, and X is the mass of the added groups (protons in this case). Similarly for the neighboring peak,

$$(m/z)_1 = \frac{M + (n_2 + 1)X}{n_2 + 1}$$

We now have two unknowns (M and n_2) and two equations. We can solve first for n_2 and then for M:

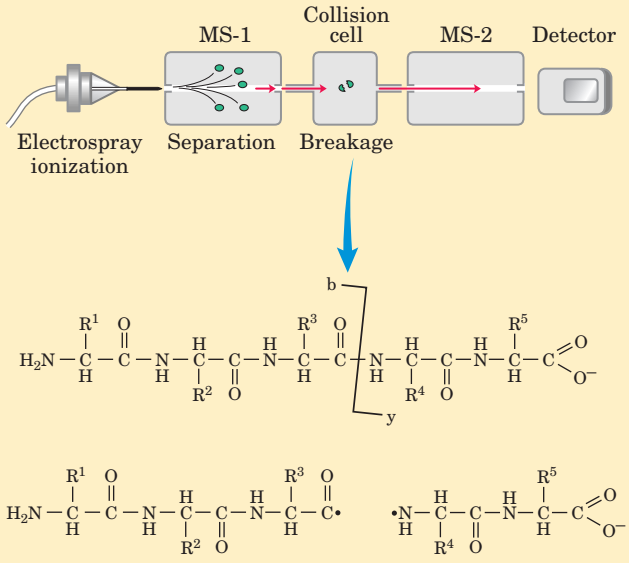
$$n_2 = \frac{(m/z)_2 - X}{(m/z)_2 - (m/z)_1}$$

$$M = n_2\,[(m/z)_2 - X]$$

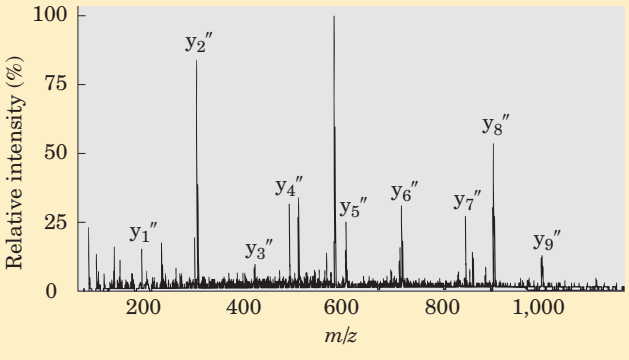
This calculation using the m/z values for any two peaks in a spectrum such as that shown in Figure 1b usually provides the mass of the protein (in this case, aerolysin k; 47,342 Da) with an error of only ±0.01%. Generating several sets of peaks, repeating the calculation, and averaging the results generally provides an even more accurate value for M. Computer algorithms can transform the m/z spectrum into a single peak that also provides a very accurate mass measurement (Fig. 1b, inset).

Mass spectrometry can also be used to sequence short stretches of polypeptide, an application that has emerged as an invaluable tool for quickly identifying unknown proteins. Sequence information is extracted using a technique called **tandem MS,** or **MS/MS.** A solution containing the protein under investigation is first treated with a protease or chemical reagent to hydrolyze it to a mixture of shorter peptides. The mixture is then injected into a device that is essentially two mass spectrometers in tandem (Fig. 2a, top). In the first, the peptide mixture is sorted and the ionized fragments are manipulated so that only one of the several types of peptides produced by cleavage emerges at the other end. The sample of the selected

(continued on next page)



(a)



(b)

FIGURE 2 Obtaining protein sequence information with tandem MS. (a) After proteolytic hydrolysis, a protein solution is injected into a mass spectrometer (MS-1). The different peptides are sorted so that only one type is selected for further analysis. The selected peptide is further fragmented in a chamber between the two mass spectrometers, and m/z for each fragment is measured in the second mass spectrometer (MS-2). Many of the ions generated during this second fragmentation result from breakage of the peptide bond, as shown. These are called b-type or y-type ions, depending on whether the charge is retained on the amino- or carboxyl-terminal side, respectively. **(b)** A typical spectrum with peaks representing the peptide fragments generated from a sample of one small peptide (10 residues). The labeled peaks are y-type ions. The large peak next to y_5'' is a doubly charged ion and is not part of the y set. The successive peaks differ by the mass of a particular amino acid in the original peptide. In this case, the deduced sequence was Phe–Pro–Gly–Gln–(Ile/Leu)–Asn–Ala–Asp–(Ile/Leu)–Arg. Note the ambiguity about Ile and Leu residues, because they have the same molecular mass. In this example, the set of peaks derived from y-type ions predominates, and the spectrum is greatly simplified as a result. This is because an Arg residue occurs at the carboxyl terminus of the peptide, and most of the positive charges are retained on this residue.

BOX 3–2 WORKING IN BIOCHEMISTRY *(continued from previous page)*

peptide, each molecule of which has a charge somewhere along its length, then travels through a vacuum chamber between the two mass spectrometers. In this collision cell, the peptide is further fragmented by high-energy impact with a "collision gas," a small amount of a noble gas such as helium or argon that is bled into the vacuum chamber. This procedure is designed to fragment many of the peptide molecules in the sample, with each individual peptide broken in only one place, on average. Most breaks occur at peptide bonds. This fragmentation does not involve the addition of water (it is done in a near-vacuum), so the products may include molecular ion radicals such as carbonyl radicals (Fig. 2a, bottom). The charge on the original peptide is retained on one of the fragments generated from it.

The second mass spectrometer then measures the m/z ratios of all the charged fragments (uncharged fragments are not detected). This generates one or more sets of peaks. A given set of peaks (Fig. 2b) consists of all the charged fragments that were generated by breaking the same type of bond (but at different points in the peptide) and are derived from the same side of the bond breakage, either the carboxyl- or amino-terminal side. Each successive peak in a given set has one less amino acid than the peak before. The difference in mass from peak to peak identifies the amino acid that was lost in each case, thus revealing the sequence of the peptide. The only ambiguities involve leucine and isoleucine, which have the same mass.

The charge on the peptide can be retained on either the carboxyl- or amino-terminal fragment, and bonds other than the peptide bond can be broken in the fragmentation process, with the result that multiple sets of peaks are usually generated. The two most prominent sets generally consist of charged fragments derived from breakage of the peptide bonds. The set consisting of the carboxyl-terminal fragments can be unambiguously distinguished from that consisting of the amino-terminal fragments. Because the bond breaks generated between the spectrometers (in the collision cell) do not yield full carboxyl and amino groups at the sites of the breaks, the only intact α-amino and α-carboxyl groups on the peptide fragments are those at the very ends (Fig. 2a). The two sets of fragments can thereby be identified by the resulting slight differences in mass. The amino acid sequence derived from one set can be confirmed by the other, improving the confidence in the sequence information obtained.

Even a short sequence is often enough to permit unambiguous association of a protein with its gene, if the gene sequence is known. Sequencing by mass spectrometry cannot replace the Edman degradation procedure for the sequencing of long polypeptides, but it is ideal for proteomics research aimed at cataloging the hundreds of cellular proteins that might be separated on a two-dimensional gel. In the coming decades, detailed genomic sequence data will be available from hundreds, eventually thousands, of organisms. The ability to rapidly associate proteins with genes using mass spectrometry will greatly facilitate the exploitation of this extraordinary information resource.

Small Peptides and Proteins Can Be Chemically Synthesized

Many peptides are potentially useful as pharmacologic agents, and their production is of considerable commercial importance. There are three ways to obtain a peptide: (1) purification from tissue, a task often made difficult by the vanishingly low concentrations of some peptides; (2) genetic engineering (Chapter 9); or (3) direct chemical synthesis. Powerful techniques now make direct chemical synthesis an attractive option in many cases. In addition to commercial applications, the synthesis of specific peptide portions of larger proteins is an increasingly important tool for the study of protein structure and function.

The complexity of proteins makes the traditional synthetic approaches of organic chemistry impractical for peptides with more than four or five amino acid residues. One problem is the difficulty of purifying the product after each step.

The major breakthrough in this technology was provided by R. Bruce Merrifield in 1962. His innovation involved synthesizing a peptide while keeping it attached at one end to a solid support. The support is an insoluble polymer (resin) contained within a column, similar to that used for chromatographic procedures. The peptide is built up on this support one amino acid at a time using a standard set of reactions in a repeating cycle (Fig. 3–29). At each successive step in the cycle, protective chemical groups block unwanted reactions.

The technology for chemical peptide synthesis is now automated. As in the sequencing reactions already considered, the most important limitation of the process is the efficiency of each chemical cycle, as can be seen by calculating the overall yields of peptides of various

lengths when the yield for addition of each new amino acid is 96.0% versus 99.8% (Table 3–8). Incomplete reaction at one stage can lead to formation of an impurity (in the form of a shorter peptide) in the next. The chemistry has been optimized to permit the synthesis of proteins of 100 amino acid residues in a few days in reasonable yield. A very similar approach is used to synthesize nucleic acids (see Fig. 8–38). It is worth noting that this technology, impressive as it is, still pales when compared with biological processes. The same

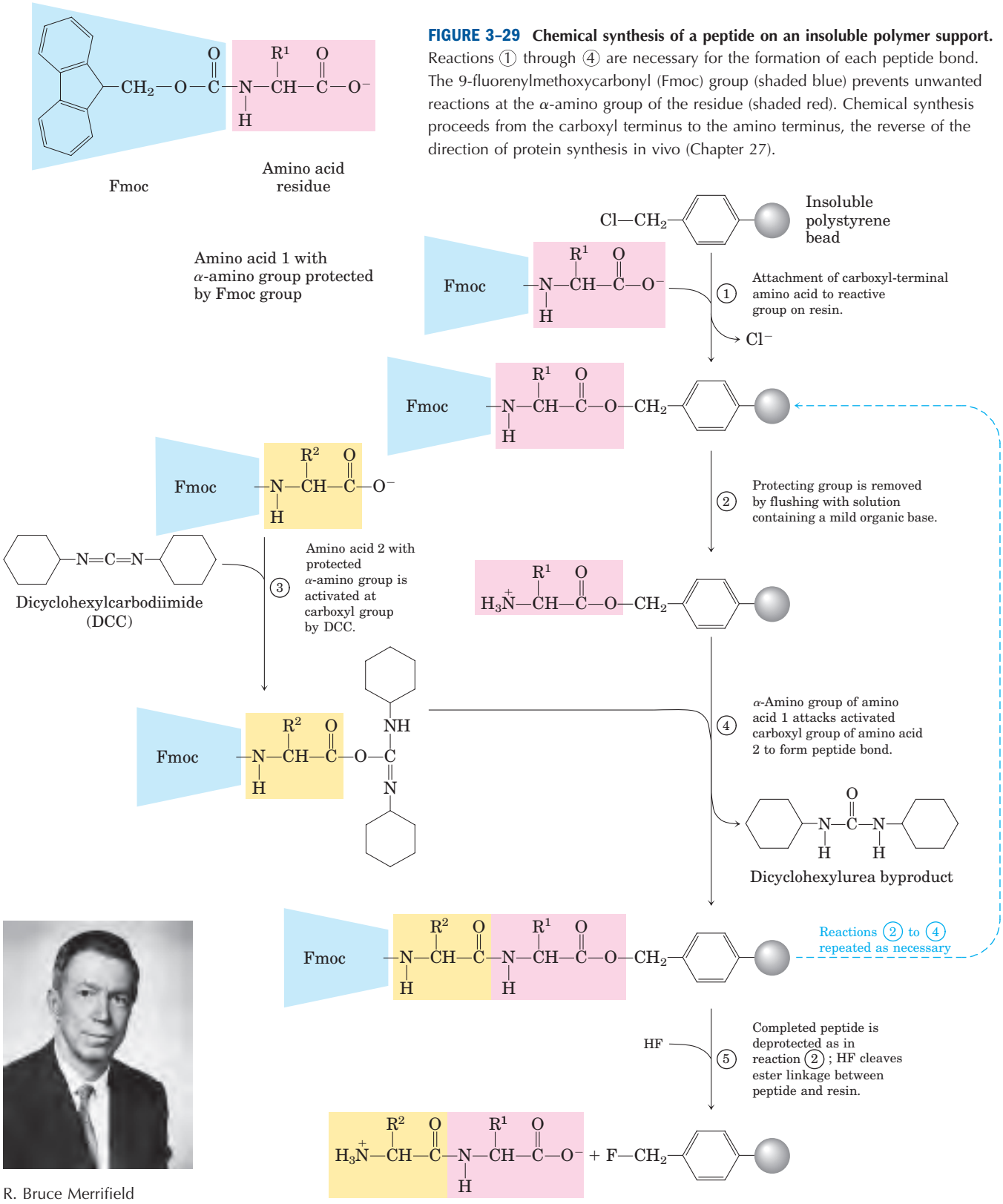


FIGURE 3–29 **Chemical synthesis of a peptide on an insoluble polymer support.** Reactions ① through ④ are necessary for the formation of each peptide bond. The 9-fluorenylmethoxycarbonyl (Fmoc) group (shaded blue) prevents unwanted reactions at the α-amino group of the residue (shaded red). Chemical synthesis proceeds from the carboxyl terminus to the amino terminus, the reverse of the direction of protein synthesis in vivo (Chapter 27).

R. Bruce Merrifield

TABLE 3–8 Effect of Stepwise Yield on Overall Yield in Peptide Synthesis

Number of residues in the final polypeptide	*Overall yield of final peptide (%) when the yield of each step is:*	
	96.0%	*99.8%*
11	66	98
21	44	96
31	29	94
51	13	90
100	1.7	82

100-amino-acid protein would be synthesized with exquisite fidelity in about 5 seconds in a bacterial cell.

A variety of new methods for the efficient ligation (joining together) of peptides has made possible the assembly of synthetic peptides into larger proteins. With these methods, novel forms of proteins can be created with precisely positioned chemical groups, including those that might not normally be found in a cellular protein. These novel forms provide new ways to test theories of enzyme catalysis, to create proteins with new chemical properties, and to design protein sequences that will fold into particular structures. This last application provides the ultimate test of our increasing ability to relate the primary structure of a peptide to the three-dimensional structure that it takes up in solution.

Amino Acid Sequences Provide Important Biochemical Information

Knowledge of the sequence of amino acids in a protein can offer insights into its three-dimensional structure and its function, cellular location, and evolution. Most of these insights are derived by searching for similarities with other known sequences. Thousands of sequences are known and available in databases accessible through the Internet. A comparison of a newly obtained sequence with this large bank of stored sequences often reveals relationships both surprising and enlightening.

Exactly how the amino acid sequence determines three-dimensional structure is not understood in detail, nor can we always predict function from sequence. However, protein families that have some shared structural or functional features can be readily identified on the basis of amino acid sequence similarities. Individual proteins are assigned to families based on the degree of similarity in amino acid sequence. Members of a family are usually identical across 25% or more of their sequences, and proteins in these families generally share at least some structural and functional characteristics. Some families are defined, however, by identities involving only a few amino acid residues that are critical to a certain function. A number of similar substructures (to be defined in Chapter 4 as "domains") occur in many functionally unrelated proteins. These domains often fold into structural configurations that have an unusual degree of stability or that are specialized for a certain environment. Evolutionary relationships can also be inferred from the structural and functional similarities within protein families.

Certain amino acid sequences serve as signals that determine the cellular location, chemical modification, and half-life of a protein. Special signal sequences, usually at the amino terminus, are used to target certain proteins for export from the cell; other proteins are targeted for distribution to the nucleus, the cell surface, the cytosol, and other cellular locations. Other sequences act as attachment sites for prosthetic groups, such as sugar groups in glycoproteins and lipids in lipoproteins. Some of these signals are well characterized and are easily recognized in the sequence of a newly characterized protein (Chapter 27).

SUMMARY 3.4 The Covalent Structure of Proteins

- Differences in protein function result from differences in amino acid composition and sequence. Some variations in sequence are possible for a particular protein, with little or no effect on function.
- Amino acid sequences are deduced by fragmenting polypeptides into smaller peptides using reagents known to cleave specific peptide bonds; determining the amino acid sequence of each fragment by the automated Edman degradation procedure; then ordering the peptide fragments by finding sequence overlaps between fragments generated by different reagents. A protein sequence can also be deduced from the nucleotide sequence of its corresponding gene in DNA.
- Short proteins and peptides (up to about 100 residues) can be chemically synthesized. The peptide is built up, one amino acid residue at a time, while remaining tethered to a solid support.

3.5 Protein Sequences and Evolution

The simple string of letters denoting the amino acid sequence of a given protein belies the wealth of information this sequence holds. As more protein sequences have become available, the development of more powerful methods for extracting information from them has become a major biochemical enterprise. Each protein's function relies on its three-dimensional structure, which

in turn is determined largely by its primary structure. Thus, the biochemical information conveyed by a protein sequence is in principle limited only by our own understanding of structural and functional principles. On a different level of inquiry, protein sequences are beginning to tell us how the proteins evolved and, ultimately, how life evolved on this planet.

Protein Sequences Can Elucidate the History of Life on Earth

The field of molecular evolution is often traced to Emile Zuckerkandl and Linus Pauling, whose work in the mid-1960s advanced the use of nucleotide and protein sequences to explore evolution. The premise is deceptively straightforward. If two organisms are closely related, the sequences of their genes and proteins should be similar. The sequences increasingly diverge as the evolutionary distance between two organisms increases. The promise of this approach began to be realized in the 1970s, when Carl Woese used ribosomal RNA sequences to define archaebacteria as a group of living organisms distinct from other bacteria and eukaryotes (see Fig. 1–4). Protein sequences offer an opportunity to greatly refine the available information. With the advent of genome projects investigating organisms from bacteria to humans, the number of available sequences is growing at an enormous rate. This information can be used to trace biological history. The challenge is in learning to read the genetic hieroglyphics.

Evolution has not taken a simple linear path. Complexities abound in any attempt to mine the evolutionary information stored in protein sequences. For a given protein, the amino acid residues essential for the activity of the protein are conserved over evolutionary time. The residues that are less important to function may vary over time—that is, one amino acid may substitute for another—and these variable residues can provide the information used to trace evolution. Amino acid substitutions are not always random, however. At some positions in the primary structure, the need to maintain protein function may mean that only particular amino acid substitutions can be tolerated. Some proteins have more variable amino acid residues than others. For these and other reasons, proteins can evolve at different rates.

Another complicating factor in tracing evolutionary history is the rare transfer of a gene or group of genes from one organism to another, a process called **lateral gene transfer.** The transferred genes may be quite similar to the genes they were derived from in the original organism, whereas most other genes in the same two organisms may be quite distantly related. An example of lateral gene transfer is the recent rapid spread of antibiotic-resistance genes in bacterial populations. The proteins derived from these transferred genes would not be good candidates for the study of bacterial evolution, because they share only a very limited evolutionary history with their "host" organisms.

The study of molecular evolution generally focuses on families of closely related proteins. In most cases, the families chosen for analysis have essential functions in cellular metabolism that must have been present in the earliest viable cells, thus greatly reducing the chance that they were introduced relatively recently by lateral gene transfer. For example, a protein called EF-1α (elongation factor 1α) is involved in the synthesis of proteins in all eukaryotes. A similar protein, EF-Tu, with the same function, is found in bacteria. Similarities in sequence and function indicate that EF-1α and EF-Tu are members of a family of proteins that share a common ancestor. The members of protein families are called **homologous proteins,** or **homologs.** The concept of a homolog can be further refined. If two proteins within a family (that is, two homologs) are present in the same species, they are referred to as **paralogs.** Homologs from different species are called **orthologs** (see Fig. 1–37). The process of tracing evolution involves first identifying suitable families of homologous proteins and then using them to reconstruct evolutionary paths.

Homologs are identified using increasingly powerful computer programs that can directly compare two or more chosen protein sequences, or can search vast databases to find the evolutionary relatives of one selected protein sequence. The electronic search process can be thought of as sliding one sequence past the other until a section with a good match is found. Within this sequence alignment, a positive score is assigned for each position where the amino acid residues in the two sequences are identical—the value of the score varying from one program to the next—to provide a measure of the quality of the alignment. The process has some complications. Sometimes the proteins being compared match well at, say, two sequence segments, and these segments are connected by less related sequences of different lengths. Thus the two matching segments cannot be aligned at the same time. To handle this, the computer program introduces "gaps" in one of the sequences to bring the matching segments into register (Fig. 3–30).



FIGURE 3–30 Aligning protein sequences with the use of gaps. Shown here is the sequence alignment of a short section of the EF-Tu protein from two well-studied bacterial species, *E. coli* and *Bacillus subtilis*. Introduction of a gap in the *B. subtilis* sequence allows a better alignment of amino acid residues on either side of the gap. Identical amino acid residues are shaded.

Of course, if a sufficient number of gaps are introduced, almost any two sequences could be brought into some sort of alignment. To avoid uninformative alignments, the programs include penalties for each gap introduced, thus lowering the overall alignment score. With electronic trial and error, the program selects the alignment with the optimal score that maximizes identical amino acid residues while minimizing the introduction of gaps.

Identical amino acids are often inadequate to identify related proteins or, more importantly, to determine how closely related the proteins are on an evolutionary time scale. A more useful analysis includes a consideration of the chemical properties of substituted amino acids. When amino acid substitutions are found within a protein family, many of the differences may be conservative—that is, an amino acid residue is replaced by a residue having similar chemical properties. For example, a Glu residue may substitute in one family member for the Asp residue found in another; both amino acids are negatively charged. Such a conservative substitution should logically garner a higher score in a sequence alignment than does a nonconservative substitution, such as the replacement of the Asp residue with a hydrophobic Phe residue.

To determine what scores to assign to the many different amino acid substitutions, Steven Henikoff and Jorja Henikoff examined the aligned sequences from a variety of different proteins. They did not analyze entire protein sequences, focusing instead on thousands of short conserved blocks where the fraction of identical amino acids was high and the alignments were thus reliable. Looking at the aligned sequence blocks, the Henikoffs analyzed the nonidentical amino acid residues within the blocks. Higher scores were given to nonidentical residues that occurred frequently than to those that appeared rarely. Even the identical residues were given scores based on how often they were replaced, such that amino acids with unique chemical properties (such as Cys and Trp) received higher scores than those more conservatively replaced (such as Asp and Glu). The result of this scoring system is a Blosum (*blo*cks *su*bstitution *m*atrix) table. The table in Figure 3–31 was generated from sequences that were identical in at least 62% of their amino acid residues, and it is thus referred to as Blosum62. Similar tables have been generated for blocks of homologous sequences that are 50% or 80% identical. When higher levels of identity are required, the most conservative amino acid substitutions can be

	A Ala	C Cys	D Asp	E Glu	F Phe	G Gly	H His	I Ile	K Lys	L Leu	M Met	N Asn	P Pro	Q Gln	R Arg	S Ser	T Thr	V Val	W Trp	Y Tyr
A	4	0	−2	−1	−2	0	−2	−1	−1	−1	−1	−2	−1	−1	−1	1	0	0	−3	−2
C		9	−3	−4	−2	−3	−3	−1	−3	−1	−1	−3	−3	−3	−3	−1	−1	−1	−2	−2
D			6	2	−3	−1	−1	−3	−1	−4	−3	1	−1	0	−2	0	−1	−3	−4	−3
E				5	−3	−2	0	−3	1	−3	−2	0	−1	2	0	0	−1	−2	−3	−2
F					6	−3	−1	0	−3	0	0	−3	−4	−3	−3	−2	−2	−1	1	3
G						6	−2	−4	−2	−4	−3	0	−2	−2	−2	0	−2	−3	−2	−3
H							8	−3	−1	−3	−2	1	−2	0	0	−1	−2	−3	−2	2
I								4	−3	2	1	−3	−3	−3	−3	−2	−1	3	−3	−1
K									5	−2	−1	0	−1	1	2	0	−1	−2	−3	−2
L										4	2	−3	−3	−2	−2	−2	−1	1	−2	−1
M											5	−2	−2	0	−1	−1	−1	1	−1	−1
N												6	−2	0	0	1	0	−3	−4	−2
P													7	−1	−2	−1	−1	−2	−4	−3
Q														5	1	0	−1	−2	−2	−1
R															5	−1	−1	−3	−3	−2
S																4	1	−2	−3	−2
T																	5	0	−2	−2
V																		4	−3	−1
W																			11	2
Y																				7

FIGURE 3–31 The Blosum62 table. This *blo*cks *su*bstitution *m*atrix was created by comparing thousands of short blocks of aligned sequences that were identical in at least 62% of their amino acid residues. The nonidentical residues were assigned scores based on how frequently they were replaced by each of the other amino acids. Each substitution contributes to the score given to a particular alignment. Positive numbers (shaded yellow) add to the score for a particular alignment; negative numbers subtract from the score. Identical residues in sequences being compared (the shaded diagonal from top left to bottom right in the matrix) receive scores based on how often they are replaced, such that amino acids with unique chemical properties (e.g., Cys and Trp) receive higher scores (9 and 11, respectively) than those more easily replaced in conservative substitutions (e.g., Asp (6) and Glu (5)). Many computer programs use Blosum62 to assign scores to new sequence alignments.

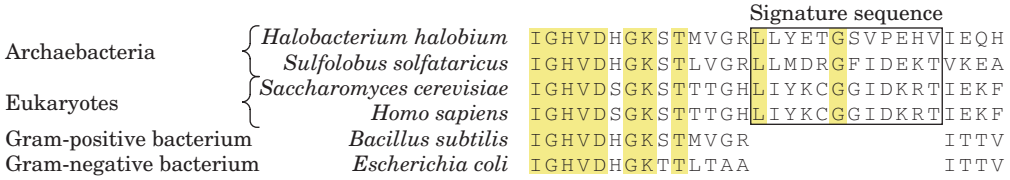


FIGURE 3–32 A signature sequence in the EF-1α/EF-Tu protein family. The signature sequence (boxed) is a 12-amino-acid insertion near the amino terminus of the sequence. Residues that align in all species are shaded yellow. Both archaebacteria and eukaryotes have the signature, although the sequences of the insertions are quite distinct for the two groups. The variation in the signature sequence reflects the significant evolutionary divergence that has occurred at this site since it first appeared in a common ancestor of both groups.

overrepresented, which limits the usefulness of the matrix in identifying homologs that are somewhat distantly related. Tests have shown that the Blosum62 table provides the most reliable alignments over a wide range of protein families, and it is the default table in many sequence alignment programs.

For most efforts to find homologies and explore evolutionary relationships, protein sequences (derived either directly from protein sequencing or from the sequencing of the DNA encoding the protein) are superior to nongenic nucleic acid sequences (those that do not encode a protein or functional RNA). For a nucleic acid, with its four different types of residues, random alignment of nonhomologous sequences will generally yield matches for at least 25% of the positions. Introduction of a few gaps can often increase the fraction of matched residues to 40% or more, and the probability of chance alignment of unrelated sequences becomes quite high. The 20 different amino acid residues in proteins greatly lower the probability of uninformative chance alignments of this type.

The programs used to generate a sequence alignment are complemented by methods that test the reliability of the alignments. A common computerized test is to shuffle the amino acid sequence of one of the proteins being compared to produce a random sequence, then instruct the program to align the shuffled sequence with the other, unshuffled one. Scores are assigned to the new alignment, and the shuffling and alignment process is repeated many times. The original alignment, before shuffling, should have a score significantly higher than any of those within the distribution of scores generated by the random alignments; this increases the confidence that the sequence alignment has identified a pair of homologs. Note that the *absence* of a significant alignment score does not necessarily mean that no evolutionary relationship exists between two proteins. As we shall see in Chapter 4, three-dimensional structural similarities sometimes reveal evolutionary relationships where sequence homology has been wiped away by time.

Using a protein family to explore evolution requires the identification of family members with similar molecular functions in the widest possible range of organisms. Information from the family can then be used to trace the evolution of those organisms. By analyzing the sequence divergence in selected protein families, investigators can segregate organisms into classes based on their evolutionary relationships. This information must be reconciled with more classical examinations of the physiology and biochemistry of the organisms.

Certain segments of a protein sequence may be found in the organisms of one taxonomic group but not in other groups; these segments can be used as **signature sequences** for the group in which they are found. An example of a signature sequence is an insertion of 12 amino acids near the amino terminus of the EF-1α/EF-Tu proteins in all archaebacteria and eukaryotes but not in other types of bacteria (Fig. 3–32). The signature is one of many biochemical clues that can help establish the evolutionary relatedness of eukaryotes and archaebacteria. For example, the major taxa of bacteria can be distinguished by signature sequences in several different proteins. The β and γ proteobacteria have signature sequences in the Hsp70 and DNA gyrase protein families (families of proteins involved in protein folding and DNA replication, respectively) that are not present in any other bacteria, including the other proteobacteria. The other types of proteobacteria (α, δ, ε), along with the β and γ proteobacteria, have a separate Hsp70 signature sequence and a signature in alanyl-tRNA synthetase (an enzyme of protein synthesis) that are not present in other bacteria. The appearance of unique signatures in the β and γ proteobacteria suggests the α, δ, and ε proteobacteria arose before their β and γ cousins.

By considering the entire sequence of a protein, researchers can now construct more elaborate evolutionary trees with many species in each taxonomic group. Figure 3–33 presents one such tree for bacteria, based on sequence divergence in the protein GroEL (a protein present in all bacteria that assists in the proper folding of proteins). The tree can be refined by basing it on the sequences of multiple proteins and by supplementing the sequence information with data on the unique biochemical and physiological properties of each species. There are many methods for generating trees, each with its own advantages and shortcomings, and

many ways to represent the resulting evolutionary relationships. In Figure 3–33, the free end points of lines are called "external nodes"; each represents an extant species, and each is so labeled. The points where two lines come together, the "internal nodes," represent extinct ancestor species. In most representations (including Fig. 3–33), the lengths of the lines connecting the nodes are proportional to the number of amino acid substitutions separating one species from another. If we trace two extant species to a common internal node (representing the common ancestor of the two species), the length of the branch connecting each external node to the internal node represents the number of amino acid substitutions separating one extant species from this ancestor. The sum of the lengths of all the line segments that connect an extant species to another extant species through a common ancestor reflects the number of substitutions separating the two extant species. To determine how much time was needed for the various species to diverge, the tree must be calibrated by comparing it with information from the fossil record and other sources.

As more sequence information is made available in databases, we can generate evolutionary trees based on a variety of different proteins. Some proteins evolve faster than others, or change faster within one group of species than another. A large protein, with many variable amino acid residues, may exhibit a few differences between two closely related species. Another, smaller protein may be identical in the same two species. For many reasons, some details of an evolutionary tree based on the sequences of one protein may differ from those of a tree based on the sequences of another protein. Increasingly sophisticated analyses using the sequences of many different proteins can provide an exquisitely detailed and accurate picture of evolutionary relationships. The story is a work in progress, and the questions being asked and answered are fundamental to how humans view themselves and the world around them. The field of molecular evolution promises to be among the most vibrant of the scientific frontiers in the twenty-first century.

SUMMARY 3.5 Protein Sequences and Evolution

- Protein sequences are a rich source of information about protein structure and function, as well as the evolution of life on this planet. Sophisticated methods are being developed to trace evolution by analyzing the resultant slow changes in the amino acid sequences of homologous proteins.

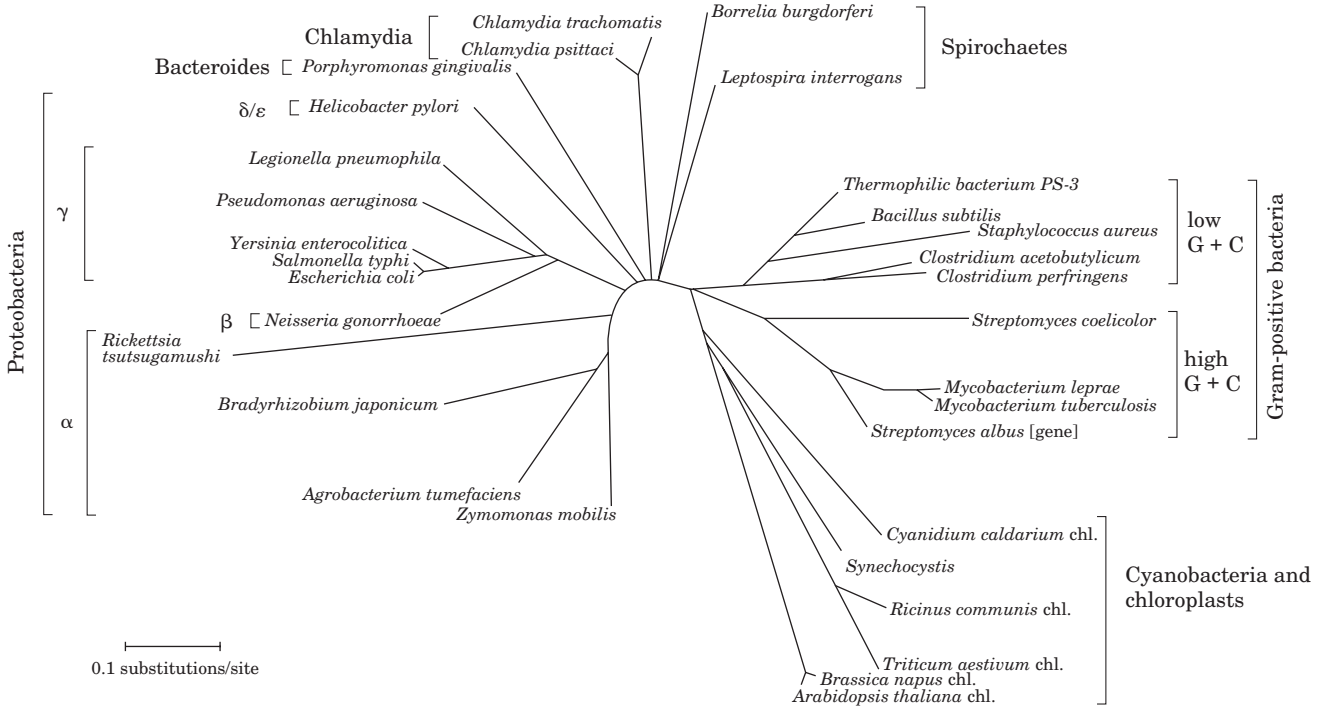


FIGURE 3–33 Evolutionary tree derived from amino acid sequence comparisons. A bacterial evolutionary tree, based on the sequence divergence observed in the GroEL family of proteins. Also included in this tree (lower right) are the chloroplasts (chl.) of some nonbacterial species.

Key Terms

Terms in bold are defined in the glossary.

amino acids 75
R group 76
chiral center 76
enantiomers 76
absolute configuration 77
D, L system 77
polarity 78
zwitterion 81
absorbance, *A* 82
isoelectric pH (isoelectric point, pI) 84
peptide 85
protein 85
peptide bond 85
oligopeptide 85
polypeptide 85
oligomeric protein 87
protomer 87
conjugated protein 88
prosthetic group 88
primary structure 88
secondary structure 88
tertiary structure 88
quaternary structure 88
crude extract 89
fractionation 89
dialysis 89
column chromatography 89
high-performance liquid chromatography (HPLC) 90
electrophoresis 92
sodium dodecyl sulfate (SDS) 92
isoelectric focusing 93
Edman degradation 98
proteases 99
proteome 101
lateral gene transfer 107
homologous proteins 107
homolog 107
paralog 107
ortholog 107
signature sequence 109

Further Reading

Amino Acids

Dougherty, D.A. (2000) Unnatural amino acids as probes of protein structure and function. *Curr. Opin. Chem. Biol.* **4,** 645–652.

Greenstein, J.P. & Winitz, M. (1961) *Chemistry of the Amino Acids,* 3 Vols, John Wiley & Sons, New York.

Kreil, G. (1997) D-Amino acids in animal peptides. *Annu. Rev. Biochem.* **66,** 337–345.
An update on the occurrence of these unusual stereoisomers of amino acids.

Meister, A. (1965) *Biochemistry of the Amino Acids,* 2nd edn, Vols 1 and 2, Academic Press, Inc., New York.
Encyclopedic treatment of the properties, occurrence, and metabolism of amino acids.

Peptides and Proteins

Creighton, T.E. (1992) *Proteins: Structures and Molecular Properties,* 2nd edn, W. H. Freeman and Company, New York.
Very useful general source.

Working with Proteins

Dunn, M.J. & Corbett, J.M. (1996) Two-dimensional polyacrylamide gel electrophoresis. *Methods Enzymol.* **271,** 177–203.
A detailed description of the technology.

Kornberg, A. (1990) Why purify enzymes? *Methods Enzymol.* **182,** 1–5.
The critical role of classical biochemical methods in a new age.

Scopes, R.K. (1994) *Protein Purification: Principles and Practice,* 3rd edn, Springer-Verlag, New York.
A good source for more complete descriptions of the principles underlying chromatography and other methods.

Covalent Structure of Proteins

Andersson, L., Blomberg, L., Flegel, M., Lepsa, L., Nilsson, B., & Verlander, M. (2000) Large-scale synthesis of peptides. *Biopolymers* **55,** 227–250.
A discussion of approaches used to manufacture peptides as pharmaceuticals.

Dell, A. & Morris, H.R. (2001) Glycoprotein structure determination by mass spectrometry. *Science* **291,** 2351–2356.
Glycoproteins can be complex; mass spectrometry is a method of choice for sorting things out.

Dongre, A.R., Eng, J.K., & Yates, J.R. III (1997) Emerging tandem-mass-spectrometry techniques for the rapid identification of proteins. *Trends Biotechnol.* **15,** 418–425.
A detailed description of mass spectrometry methods.

Gygi, S.P. & Aebersold, R. (2000) Mass spectrometry and proteomics. *Curr. Opin. Chem. Biol.* **4,** 489–494.
Uses of mass spectrometry to identify and study cellular proteins.

Koonin, E.V., Tatusov, R.L., & Galperin, M.Y. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* **8,** 355–363.
A good discussion about the possible uses of the tremendous amount of protein sequence information becoming available.

Mann, M. & Wilm, M. (1995) Electrospray mass spectrometry for protein characterization. *Trends Biochem. Sci.* **20,** 219–224.
An approachable summary of this technique for beginners.

Mayo, K.H. (2000) Recent advances in the design and construction of synthetic peptides: for the love of basics or just for the technology of it. *Trends Biotechnol.* **18,** 212–217.

Miranda, L.P. & Alewood, P.F. (2000) Challenges for protein chemical synthesis in the 21st century: bridging genomics and proteomics. *Biopolymers* **55,** 217–226.

This and the Mayo article describe how to make peptides and splice them together to address a wide range of problems in protein biochemistry.

Sanger, F. (1988) Sequences, sequences, sequences. *Annu. Rev. Biochem.* **57,** 1–28.

A nice historical account of the development of sequencing methods.

Protein Sequences and Evolution

Gupta, R.S. (1998) Protein phylogenies and signal sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiol. Mol. Biol. Rev.* **62,** 1435–1491.

An almost encyclopedic but very readable report of how protein sequences are used to explore evolution, introducing many interesting ideas and supporting them with detailed sequence comparisons.

Li, W.-H. & Graur, D. (2000) *Fundamentals of Molecular Evolution*, 2nd edn, Sinauer Associates, Inc., Sunderland, MA.

A very readable text describing methods used to analyze protein and nucleic acid sequences. Chapter 5 provides one of the best available descriptions of how evolutionary trees are constructed from sequence data.

Rokas, A., Williams, B.L., King, N., & Carroll, S.B. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425,** 798–804.

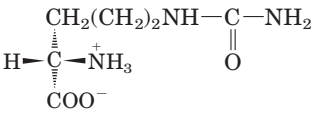
How sequence comparisons of multiple proteins can yield accurate evolutionary information.

Zuckerkandl, E. & Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* **8,** 357–366.

Considered by many the founding paper in the field of molecular evolution.

Problems

1. Absolute Configuration of Citrulline The citrulline isolated from watermelons has the structure shown below. Is it a D- or L-amino acid? Explain.



2. Relationship between the Titration Curve and the Acid-Base Properties of Glycine A 100 mL solution of 0.1 M glycine at pH 1.72 was titrated with 2 M NaOH solution. The pH was monitored and the results were plotted on a graph, as shown at right. The key points in the titration are designated I to V. For each of the statements (a) to (o), *identify* the appropriate key point in the titration and *justify* your choice.

(a) Glycine is present predominantly as the species $^{+}H_3N—CH_2—COOH$.

(b) The *average* net charge of glycine is $+\frac{1}{2}$.

(c) Half of the amino groups are ionized.

(d) The pH is equal to the pK_a of the carboxyl group.

(e) The pH is equal to the pK_a of the protonated amino group.

(f) Glycine has its maximum buffering capacity.

(g) The *average* net charge of glycine is zero.

(h) The carboxyl group has been completely titrated (first equivalence point).

(i) Glycine is completely titrated (second equivalence point).

(j) The predominant species is $^{+}H_3N—CH_2—COO^{-}$.

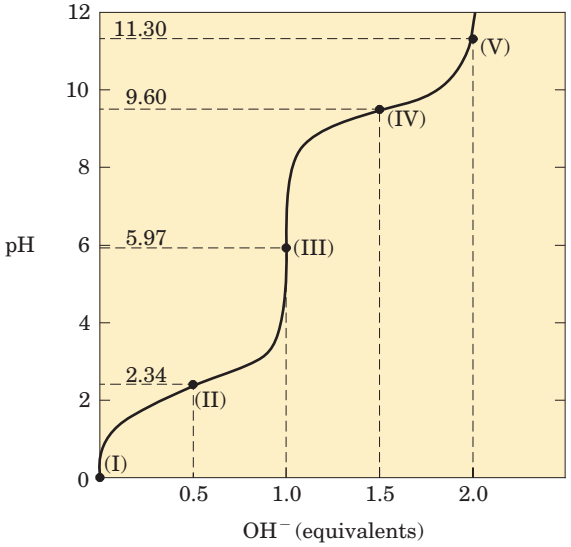
(k) The *average* net charge of glycine is -1.

(l) Glycine is present predominantly as a 50:50 mixture of $^{+}H_3N—CH_2—COOH$ and $^{+}H_3N—CH_2—COO^{-}$.

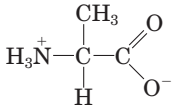
(m) This is the isoelectric point.

(n) This is the end of the titration.

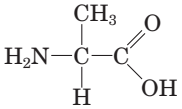
(o) These are the *worst* pH regions for buffering power.



3. How Much Alanine Is Present as the Completely Uncharged Species? At a pH equal to the isoelectric point of alanine, the *net* charge on alanine is zero. Two structures can be drawn that have a net charge of zero, but the predominant form of alanine at its pI is zwitterionic.



Zwitterionic



Uncharged

(a) Why is alanine predominantly zwitterionic rather than completely uncharged at its pI?

(b) What fraction of alanine is in the completely uncharged form at its pI? Justify your assumptions.

4. Ionization State of Amino Acids Each ionizable group of an amino acid can exist in one of two states, charged or neutral. The electric charge on the functional group is determined by the relationship between its pK_a and the pH of the solution. This relationship is described by the Henderson-Hasselbalch equation.

(a) Histidine has three ionizable functional groups. Write the equilibrium equations for its three ionizations and assign the proper pK_a for each ionization. Draw the structure of histidine in each ionization state. What is the net charge on the histidine molecule in each ionization state?

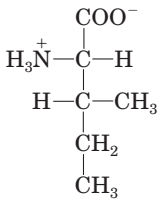
(b) Draw the structures of the predominant ionization state of histidine at pH 1, 4, 8, and 12. Note that the ionization state can be approximated by treating each ionizable group independently.

(c) What is the net charge of histidine at pH 1, 4, 8, and 12? For each pH, will histidine migrate toward the anode (+) or cathode (−) when placed in an electric field?

5. Separation of Amino Acids by Ion-Exchange Chromatography Mixtures of amino acids are analyzed by first separating the mixture into its components through ion-exchange chromatography. Amino acids placed on a cation-exchange resin containing sulfonate groups (see Fig. 3–18a) flow down the column at different rates because of two factors that influence their movement: (1) ionic attraction between the $—SO_3^-$ residues on the column and positively charged functional groups on the amino acids, and (2) hydrophobic interactions between amino acid side chains and the strongly hydrophobic backbone of the polystyrene resin. For each pair of amino acids listed, determine which will be eluted first from an ion-exchange column using a pH 7.0 buffer.

(a) Asp and Lys
(b) Arg and Met
(c) Glu and Val
(d) Gly and Leu
(e) Ser and Ala

6. Naming the Stereoisomers of Isoleucine The structure of the amino acid isoleucine is



(a) How many chiral centers does it have?
(b) How many optical isomers?
(c) Draw perspective formulas for all the optical isomers of isoleucine.

7. Comparing the pK_a Values of Alanine and Polyalanine The titration curve of alanine shows the ionization of two functional groups with pK_a values of 2.34 and 9.69, corresponding to the ionization of the carboxyl and the protonated amino groups, respectively. The titration of di-, tri-, and larger oligopeptides of alanine also shows the ionization of only two functional groups, although the experimental pK_a values are different. The trend in pK_a values is summarized in the table.

Amino acid or peptide	pK_1	pK_2
Ala	2.34	9.69
Ala–Ala	3.12	8.30
Ala–Ala–Ala	3.39	8.03
Ala–(Ala)$_n$–Ala, $n \geq 4$	3.42	7.94

(a) Draw the structure of Ala–Ala–Ala. Identify the functional groups associated with pK_1 and pK_2.

(b) Why does the value of pK_1 *increase* with each addition of an Ala residue to the Ala oligopeptide?

(c) Why does the value of pK_2 *decrease* with each addition of an Ala residue to the Ala oligopeptide?

8. The Size of Proteins What is the approximate molecular weight of a protein with 682 amino acid residues in a single polypeptide chain?

9. The Number of Tryptophan Residues in Bovine Serum Albumin A quantitative amino acid analysis reveals that bovine serum albumin (BSA) contains 0.58% tryptophan (M_r 204) by weight.

(a) Calculate the *minimum* molecular weight of BSA (i.e., assuming there is only one tryptophan residue per protein molecule).

(b) Gel filtration of BSA gives a molecular weight estimate of 70,000. How many tryptophan residues are present in a molecule of serum albumin?

10. Net Electric Charge of Peptides A peptide has the sequence

Glu–His–Trp–Ser–Gly–Leu–Arg–Pro–Gly

(a) What is the net charge of the molecule at pH 3, 8, and 11? (Use pK_a values for side chains and terminal amino and carboxyl groups as given in Table 3–1.)

(b) Estimate the pI for this peptide.

11. Isoelectric Point of Pepsin Pepsin is the name given to several digestive enzymes secreted (as larger precursor proteins) by glands that line the stomach. These glands also secrete hydrochloric acid, which dissolves the particulate matter in food, allowing pepsin to enzymatically cleave individual protein molecules. The resulting mixture of food, HCl, and digestive enzymes is known as chyme and has a pH near 1.5. What pI would you predict for the pepsin proteins? What functional groups must be present to confer this pI on pepsin? Which amino acids in the proteins would contribute such groups?

12. The Isoelectric Point of Histones Histones are proteins found in eukaryotic cell nuclei, tightly bound to DNA, which has many phosphate groups. The pI of histones is very high, about 10.8. What amino acid residues must be present in relatively large numbers in histones? In what way do these residues contribute to the strong binding of histones to DNA?

13. Solubility of Polypeptides One method for separating polypeptides makes use of their differential solubilities. The solubility of large polypeptides in water depends upon the relative polarity of their R groups, particularly on the number of ionized groups: the more ionized groups there are, the more soluble the polypeptide. Which of each pair of the polypeptides that follow is more soluble at the indicated pH?

(a) $(Gly)_{20}$ or $(Glu)_{20}$ at pH 7.0
(b) $(Lys–Ala)_3$ or $(Phe–Met)_3$ at pH 7.0
(c) $(Ala–Ser–Gly)_5$ or $(Asn–Ser–His)_5$ at pH 6.0
(d) $(Ala–Asp–Gly)_5$ or $(Asn–Ser–His)_5$ at pH 3.0

14. Purification of an Enzyme A biochemist discovers and purifies a new enzyme, generating the purification table below.

Procedure	*Total protein (mg)*	*Activity (units)*
1. Crude extract	20,000	4,000,000
2. Precipitation (salt)	5,000	3,000,000
3. Precipitation (pH)	4,000	1,000,000
4. Ion-exchange chromatography	200	800,000
5. Affinity chromatography	50	750,000
6. Size-exclusion chromatography	45	675,000

(a) From the information given in the table, calculate the specific activity of the enzyme solution after each purification procedure.

(b) Which of the purification procedures used for this enzyme is most effective (i.e., gives the greatest relative increase in purity)?

(c) Which of the purification procedures is least effective?

(d) Is there any indication based on the results shown in the table that the enzyme after step 6 is now pure? What else could be done to estimate the purity of the enzyme preparation?

15. Sequence Determination of the Brain Peptide Leucine Enkephalin A group of peptides that influence nerve transmission in certain parts of the brain has been isolated from normal brain tissue. These peptides are known as opioids, because they bind to specific receptors that also bind opiate drugs, such as morphine and naloxone. Opioids thus mimic some of the properties of opiates. Some researchers consider these peptides to be the brain's own pain killers. Using the information below, determine the amino acid sequence of the opioid leucine enkephalin. Explain how your structure is consistent with each piece of information.

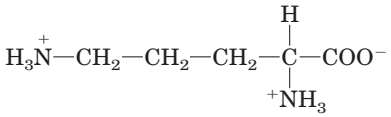
(a) Complete hydrolysis by 6 M HCl at 110 °C followed by amino acid analysis indicated the presence of Gly, Leu, Phe, and Tyr, in a 2:1:1:1 molar ratio.

(b) Treatment of the peptide with 1-fluoro-2,4-dinitrobenzene followed by complete hydrolysis and chromatography indicated the presence of the 2,4-dinitrophenyl derivative of tyrosine. No free tyrosine could be found.

(c) Complete digestion of the peptide with pepsin followed by chromatography yielded a dipeptide containing Phe and Leu, plus a tripeptide containing Tyr and Gly in a 1:2 ratio.

16. Structure of a Peptide Antibiotic from *Bacillus brevis* Extracts from the bacterium *Bacillus brevis* contain a peptide with antibiotic properties. This peptide forms complexes with metal ions and apparently disrupts ion transport across the cell membranes of other bacterial species, killing them. The structure of the peptide has been determined from the following observations.

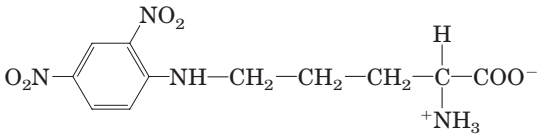
(a) Complete acid hydrolysis of the peptide followed by amino acid analysis yielded equimolar amounts of Leu, Orn, Phe, Pro, and Val. Orn is ornithine, an amino acid not present in proteins but present in some peptides. It has the structure



(b) The molecular weight of the peptide was estimated as about 1,200.

(c) The peptide failed to undergo hydrolysis when treated with the enzyme carboxypeptidase. This enzyme catalyzes the hydrolysis of the carboxyl-terminal residue of a polypeptide unless the residue is Pro or, for some reason, does not contain a free carboxyl group.

(d) Treatment of the intact peptide with 1-fluoro-2,4-dinitrobenzene, followed by complete hydrolysis and chromatography, yielded only free amino acids and the following derivative:



(Hint: Note that the 2,4-dinitrophenyl derivative involves the amino group of a side chain rather than the α-amino group.)

(e) Partial hydrolysis of the peptide followed by chromatographic separation and sequence analysis yielded the following di- and tripeptides (the amino-terminal amino acid is always at the left):

Leu–Phe Phe–Pro Orn–Leu Val–Orn

Val–Orn–Leu Phe–Pro–Val Pro–Val–Orn

Given the above information, deduce the amino acid sequence of the peptide antibiotic. Show your reasoning. When you have arrived at a structure, demonstrate that it is consistent with *each* experimental observation.

17. Efficiency in Peptide Sequencing A peptide with the primary structure Lys–Arg–Pro–Leu–Ile–Asp–Gly–Ala is sequenced by the Edman procedure. If each Edman cycle is 96% efficient, what percentage of the amino acids liberated in the fourth cycle will be leucine? Do the calculation a second time, but assume a 99% efficiency for each cycle.

18. Biochemistry Protocols: Your First Protein Purification As the newest and least experienced student in a biochemistry research lab, your first few weeks are spent washing glassware and labeling test tubes. You then graduate to making buffers and stock solutions for use in various laboratory procedures. Finally, you are given responsibility for purifying a protein. It is a citric acid cycle enzyme, citrate synthase, located in the mitochondrial matrix. Following a protocol for the purification, you proceed through the steps below. As you work, a more experienced student questions you about the rationale for each procedure. Supply the answers. (Hint: See Chapter 2 for information about osmolarity; see p. 6 for information on separation of organelles from cells.)

(a) You pick up 20 kg of beef hearts from a nearby slaughterhouse. You transport the hearts on ice, and perform

each step of the purification on ice or in a walk-in cold room. You homogenize the beef heart tissue in a high-speed blender in a medium containing 0.2 M sucrose, buffered to a pH of 7.2. *Why do you use beef heart tissue, and in such large quantity? What is the purpose of keeping the tissue cold and suspending it in 0.2 M sucrose, at pH 7.2? What happens to the tissue when it is homogenized?*

(b) You subject the resulting heart homogenate, which is dense and opaque, to a series of differential centrifugation steps. *What does this accomplish?*

(c) You proceed with the purification using the supernatant fraction that contains mostly intact mitochondria. Next you osmotically lyse the mitochondria. The lysate, which is less dense than the homogenate, but still opaque, consists primarily of mitochondrial membranes and internal mitochondrial contents. To this lysate you add ammonium sulfate, a highly soluble salt, to a specific concentration. You centrifuge the solution, decant the supernatant, and discard the pellet. To the supernatant, which is clearer than the lysate, you add *more* ammonium sulfate. Once again, you centrifuge the sample, but this time you save the pellet because it contains the protein of interest. *What is the rationale for the two-step addition of the salt?*

(d) You solubilize the ammonium sulfate pellet containing the mitochondrial proteins and dialyze it overnight against large volumes of buffered (pH 7.2) solution. *Why isn't ammonium sulfate included in the dialysis buffer? Why do you use the buffer solution instead of water?*

(e) You run the dialyzed solution over a size-exclusion chromatographic column. Following the protocol, you collect the *first* protein fraction that exits the column, and discard the rest of the fractions that elute from the column later. You detect the protein by measuring UV absorbance (at 280 nm) in the fractions. *What does the instruction to collect the first fraction tell you about the protein? Why is UV absorbance at 280 nm a good way to monitor for the presence of protein in the eluted fractions?*

(f) You place the fraction collected in (e) on a cation-exchange chromatographic column. After discarding the initial solution that exits the column (the flowthrough), you add a washing solution of higher pH to the column and collect the protein fraction that immediately elutes. *Explain what you are doing.*

(g) You run a small sample of your fraction, now very reduced in volume and quite clear (though tinged pink), on an isoelectric focusing gel. When stained, the gel shows three sharp bands. According to the protocol, the protein of interest is the one with the pI of 5.6, but you decide to do one more assay of the protein's purity. You cut out the pI 5.6 band and subject it to SDS polyacrylamide gel electrophoresis. The protein resolves as a single band. *Why were you unconvinced of the purity of the "single" protein band on your isoelectric focusing gel? What did the results of the SDS gel tell you? Why is it important to do the SDS gel electrophoresis* after *the isoelectric focusing?*