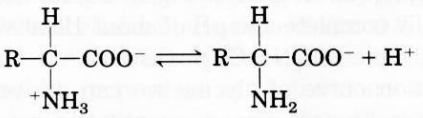
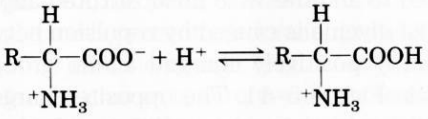
Amino Acids Can Act as Acids and Bases

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 5–9. A zwitterion can act as either an acid (proton donor):



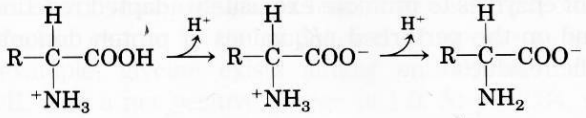
Zwitterion

or a base (proton acceptor):



Zwitterion

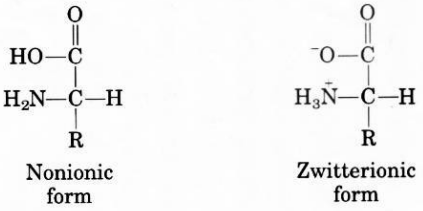
Substances having this dual nature are **amphoteric** and are often called **ampholytes** (from "amphoteric electrolytes"). A simple monoamino monocarboxylic α-amino acid, such as alanine, is a diprotic acid when fully protonated—it has two groups, the —COOH group and the $-NH_3^+$ group, that can yield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



Net charge: +1 0 −1

Amino Acids Have Characteristic Titration Curves

Acid-base titration involves the gradual addition or removal of protons (Chapter 4). Figure 5–10 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. 4–15), and can be analyzed in the same way. At very low pH, the predominant ionic species of glycine is $^+H_3N—CH_2—COOH$, the fully protonated form. 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. 4–16). For glycine, the pH at the midpoint is 2.34, thus its —COOH group has a pK_a (labeled pK_1 in Fig. 5–10) of 2.34. (Recall from Chapter 4 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

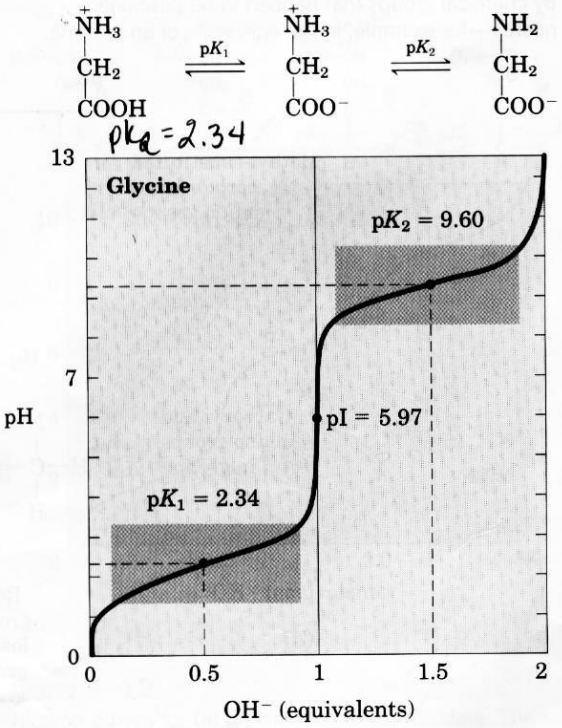


Nonionic form Zwitterionic form

figure 5–9
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.

pH = pK_a + log [ ]

figure 5–10
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.



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 (pI in Fig. 5–10) 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. 5–10) 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 $H_2N—CH_2—COO^{-}$.

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 4, 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 5–11. 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 regardless of the carboxyl group charge, 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. 5–11). 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.

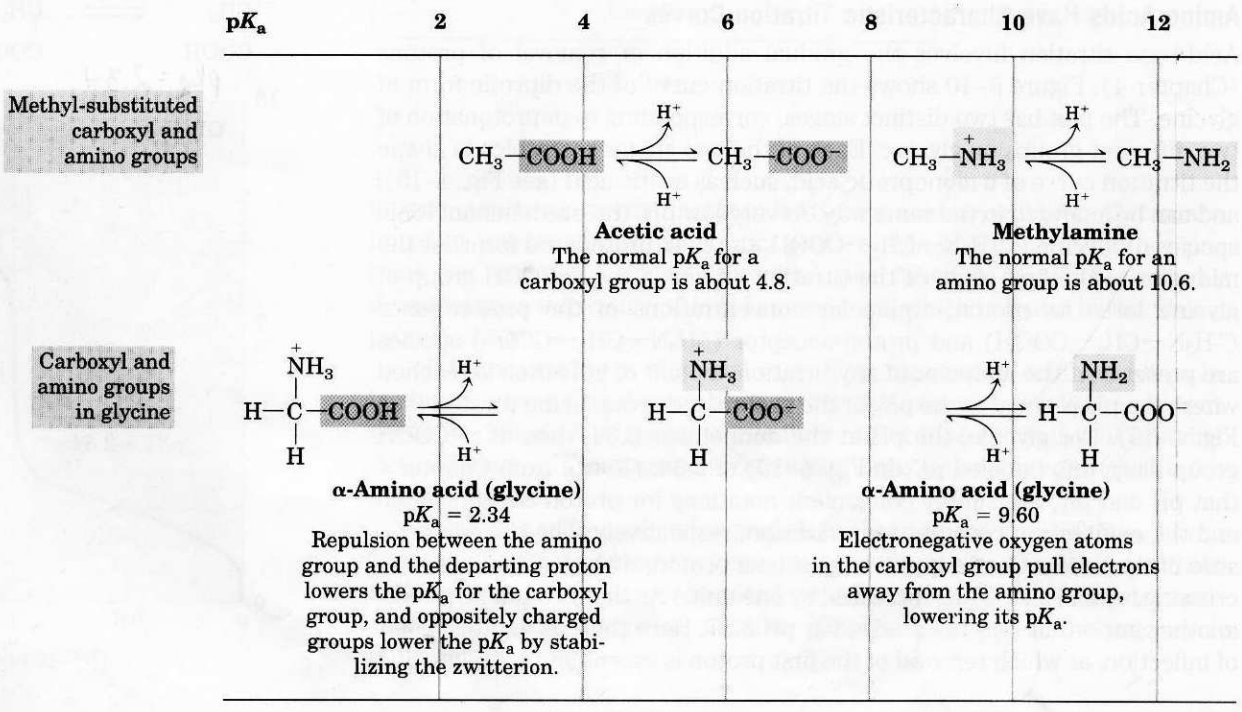


figure 5–11
Effect of the chemical environment on pK_a. The pK_a values for the ionizable groups in glycine are lower than those for simple, methyl-substituted amino and carboxyl groups. These downward perturbations of pK_a are due to intramolecular interactions. Similar effects can be caused by chemical 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 (Fig. 5–10) is that this amino acid has *two* regions of buffering power (see Fig. 4–17). One of these is the relatively flat portion of the curve, extending for approximately one 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 (Chapter 4) 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. 5–10). 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:

$$pI = \frac{1}{2}(pK_1 + pK_2) = \frac{1}{2}(2.34 + 9.60) = 5.97$$

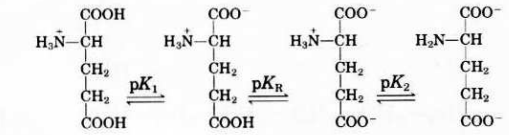
As is evident in Figure 5–10, glycine has a net negative charge at any pH above its pI and will thus move toward the positive electrode (the anode) when placed in an electric field. At any pH below its pI, glycine has a net positive charge and will move toward the negative electrode (the cathode). The 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.

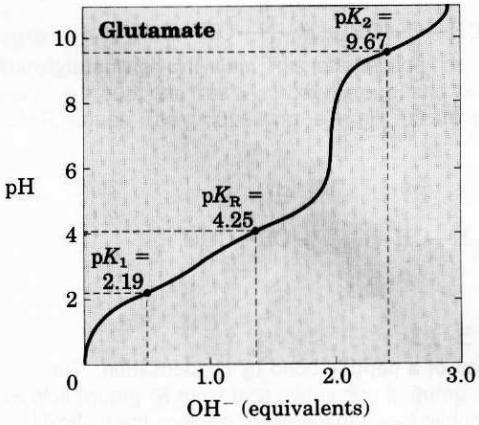
Amino Acids Differ in Their Acid-Base Properties

The shared properties of many amino acids permit some simplifying generalizations about their acid-base behaviors.

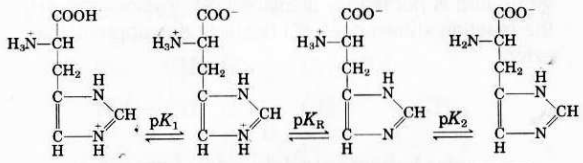
All amino acids with a single α-amino group, a single α-carboxyl group, and an R group that does not ionize have titration curves resembling that of glycine (Fig. 5–10). These amino acids have very similar, although not identical, pK_a values: pK_a of the —COOH group in the range of 1.8 to 2.4, and pK_a of the $—NH_3^+$ group in the range of 8.8 to 11.0 (Table 5–1).

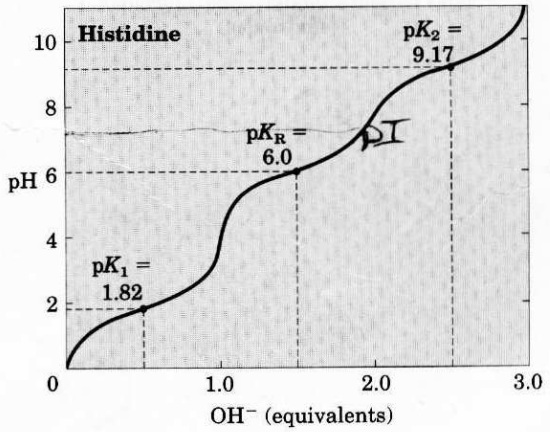
Amino acids with an ionizable R group have more complex titration curves, with *three* stages corresponding to the three possible ionization steps; thus they have three pK_a values. The additional stage for the titration of the ionizable R group merges to some extent with the other two. The titration curves for two amino acids of this type, glutamate and histidine, are shown in Figure 5–12. The isoelectric points reflect the nature of the ionizing R groups present. For example, glutamate 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 negative 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.





(a)





(b)

figure 5–12
Titration curves for **(a)** glutamate and **(b)** histidine. The pK_a of the R group is designated here as pK_R.

Another important generalization can be made about the acid-base behavior of the 20 standard amino acids. 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 intercellular fluids of most animals and bacteria. No other amino acid has an ionizable side chain with a pK_a value near enough to pH 7.0 to be an effective physiological buffer (Table 5–1).

Peptides and Proteins

We now turn to polymers of amino acids, the **peptides** and **proteins.** Biologically occurring peptides range in size from small to very large, consisting of two or three to thousands of linked amino acid residues. The focus here 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. 5–13). Peptide bond formation is an example of a condensation reaction, a common class of reaction in living cells. Under standard biochemical conditions the reaction shown in Figure 5–13 has an equilibrium that favors reactants rather than products. 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.

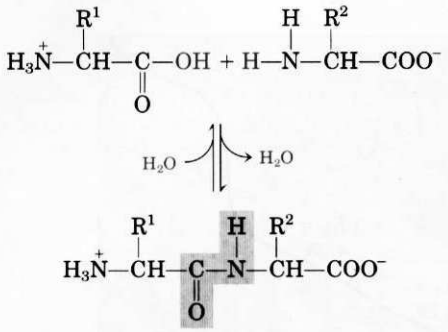


figure 5–13
Formation of a peptide bond by condensation. The α-amino group of one amino acid (with R^2 group) acts as a nucleophile (see Table 3–4) to displace the hydroxyl group of another amino acid (with R^1 group), forming a peptide bond (shaded in gray). Amino groups are good nucleophiles, but the 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.

Three amino acids can be joined by two peptide bonds to form a tripeptide; similarly, amino acids can be linked to form tetrapeptides and pentapeptides. 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.

Figure 5–14 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 a 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 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 a half-life ($t_{1/2}$) of about 7 years under most intracellular conditions.

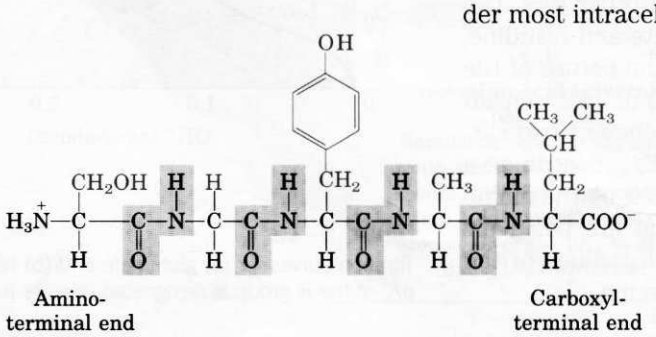


figure 5–14
The pentapeptide serylglycyltyrosylalanylleucine, or Ser–Gly–Tyr–Ala–Leu. Peptides are named beginning with the amino-terminal residue, which by convention is placed at the left. The peptide bonds are shaded in gray, the R groups are in red.

Peptides Can Be Distinguished by Their Ionization Behavior

Peptides contain only one free α-amino group and one free α-carboxyl group, one at each end of the chain (Fig. 5–15). These groups ionize as they do in free amino acids, although the ionization constants are different because the oppositely charged group is absent from the α carbon. The α-amino and α-carboxyl groups of all nonterminal amino acids are covalently joined in the form of 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 5–1), and in a peptide these contribute to the overall acid-base properties of the molecule (Fig. 5–15). 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. 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, 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 5–1 can be a useful guide to the pH range in which a given group will ionize, but they cannot be strictly applied to peptides.

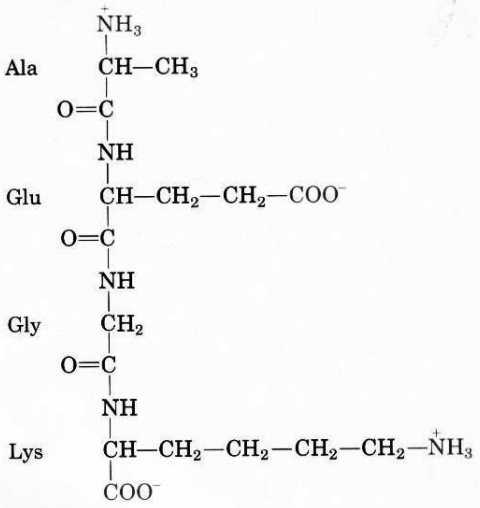
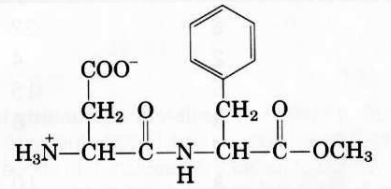


figure 5–15
Alanylglutamylglycyllysine. This tetrapeptide has one free α-amino group, one free α-carboxyl group, and two ionizable R groups. The groups ionized at pH 7.0 are in red.

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 function. Naturally occurring peptides range in length from two amino acids to many thousands of 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 5–2 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 polypeptides are much smaller than this, containing less 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

table 5–2

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 (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

noncovalently (Table 5–2). 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 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 group by dividing its molecular weight by 110. Although the average molecular weight of the 20 standard 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 5–1), the average molecular weight 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 standard amino acids almost never occur in equal amounts in a protein. Some amino acids may occur only once per molecule or not at all in a given type of protein; others may occur in large numbers. Table 5–3 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.

table 5–3

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

*Note that standard procedures for the acid hydrolysis of proteins convert Asn and Gln to Asp and Glu, respectively. In addition, Trp is destroyed. Special procedures must be employed to determine the amounts of these amino acids.

Some Proteins Contain Chemical Groups Other Than Amino Acids

Many proteins, for example the enzymes ribonuclease and chymotrypsinogen, contain only amino acid residues and no other chemical groups; 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 5–4); 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 5–4

Conjugated Proteins

Class	Prosthetic group(s)	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. 5–16). 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.**

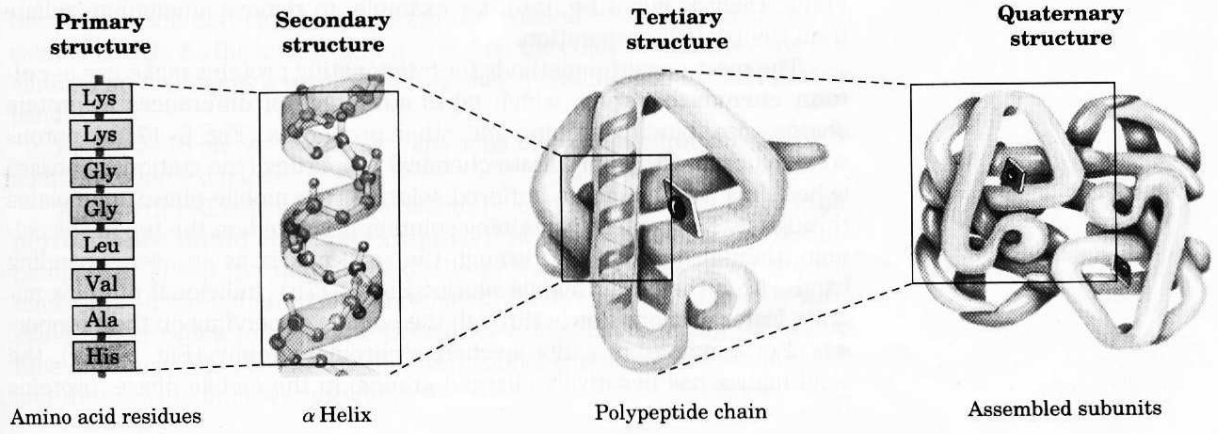


figure 5–16

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.

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 any detail it must be separated from all other proteins, and techniques must be available 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 of a protein is essential before its properties, amino acid composition, and sequence 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. For example, many proteins bind to other biomolecules with great specificity, and such proteins can be separated on the basis of their 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. 2–20).

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 some 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 amounts 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 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. 5–17). 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 is layered on the top of the column, then also percolates through the solid matrix as an ever-expanding band within the larger mobile phase (Fig. 5–17b). Individual proteins migrate faster or more slowly through the column depending on their properties. For example, in cation-exchange chromatography (Fig. 5–18a), the solid matrix has negatively charged groups. In the mobile phase, proteins

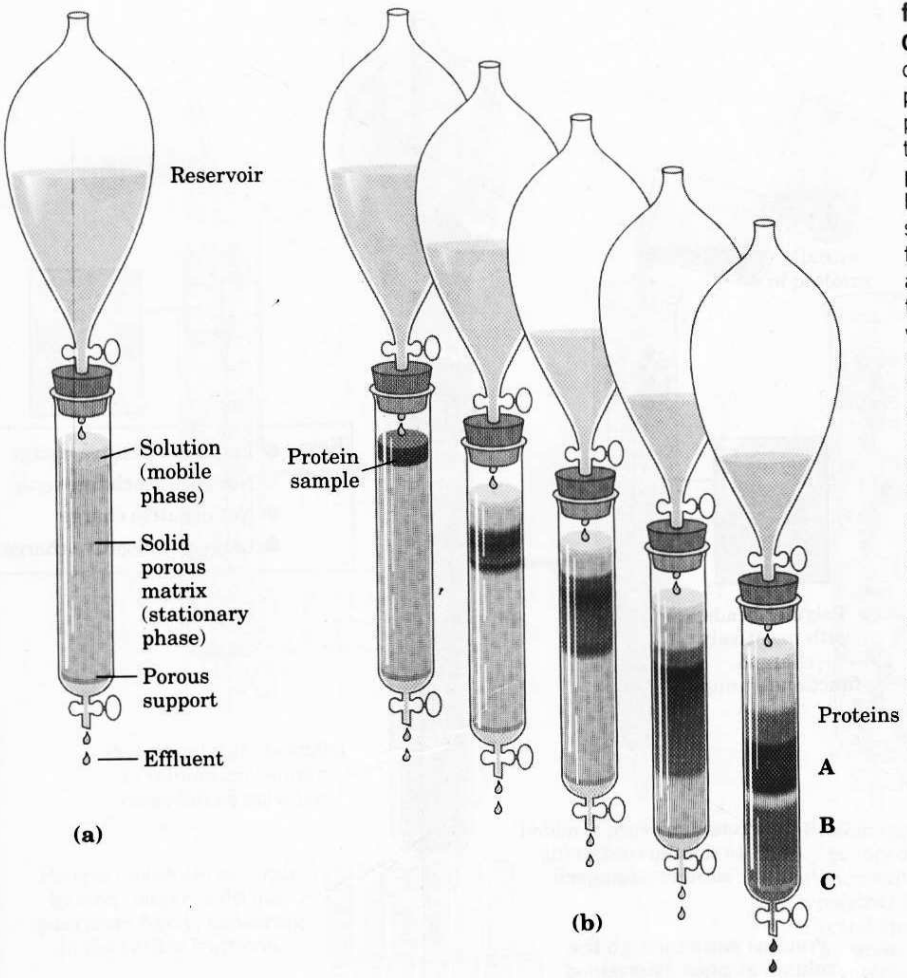


figure 5–17
Column chromatography. (a) The standard elements of a chromatographic column. A solid, porous material is supported inside a column generally made of some form of plastic. The solid material (matrix) makes up the stationary phase through which flows a solution, the mobile phase. The solution that passes out of the column at the bottom (the effluent) is constantly replaced by solution supplied from a reservoir at the top. **(b)** 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.

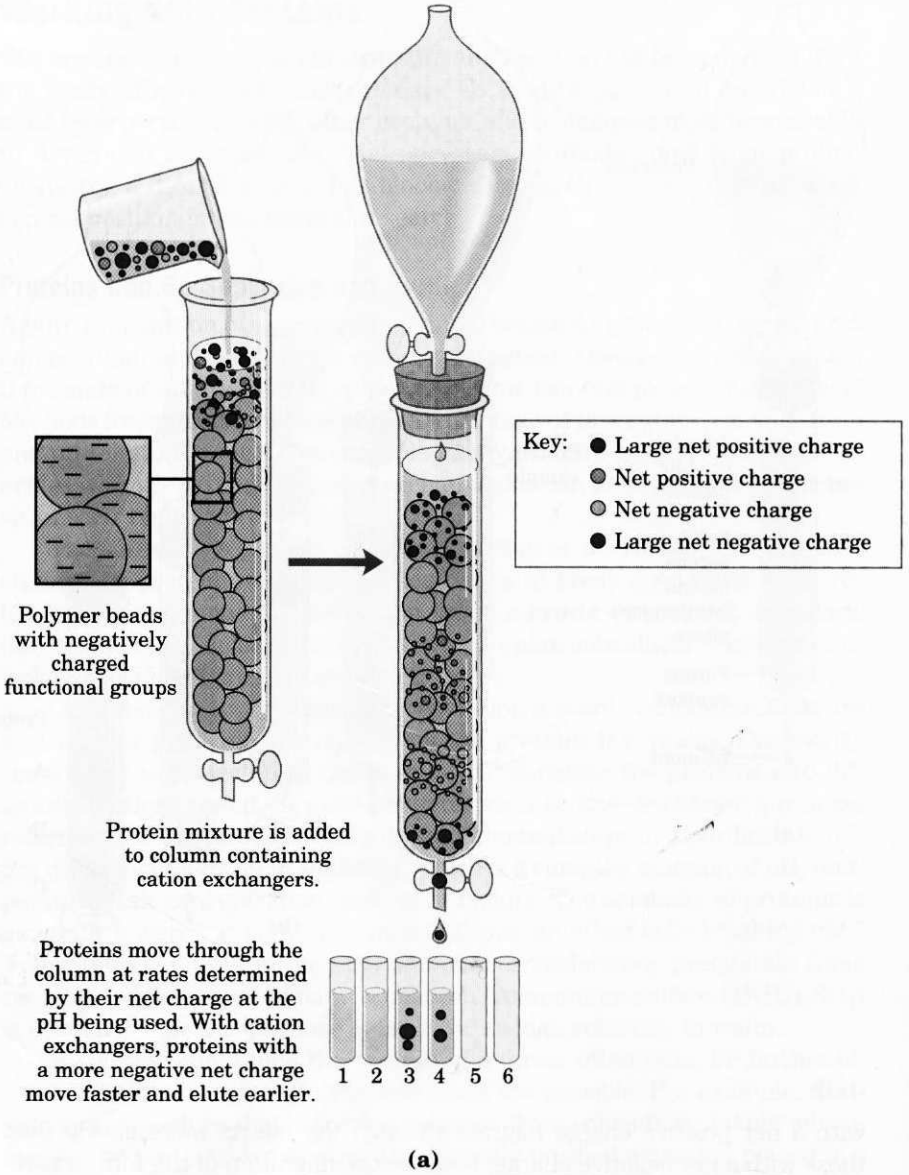
with a net positive charge migrate through the matrix more slowly than those with a net negative charge, because the migration of the former is retarded more by interaction with the stationary phase. The two types of protein can separate into two distinct bands. The expansion of the protein band in the 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. 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 5–18 shows two other variations of column chromatography in addition to ion exchange.

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.

figure 5–18

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 the more labyrinthine path they take 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.



The approach to the purification of a protein that has not been previously 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 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 number of contaminants is 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 5–5), making it feasible to use more sophisticated (and expensive) chromatographic procedures at later stages.

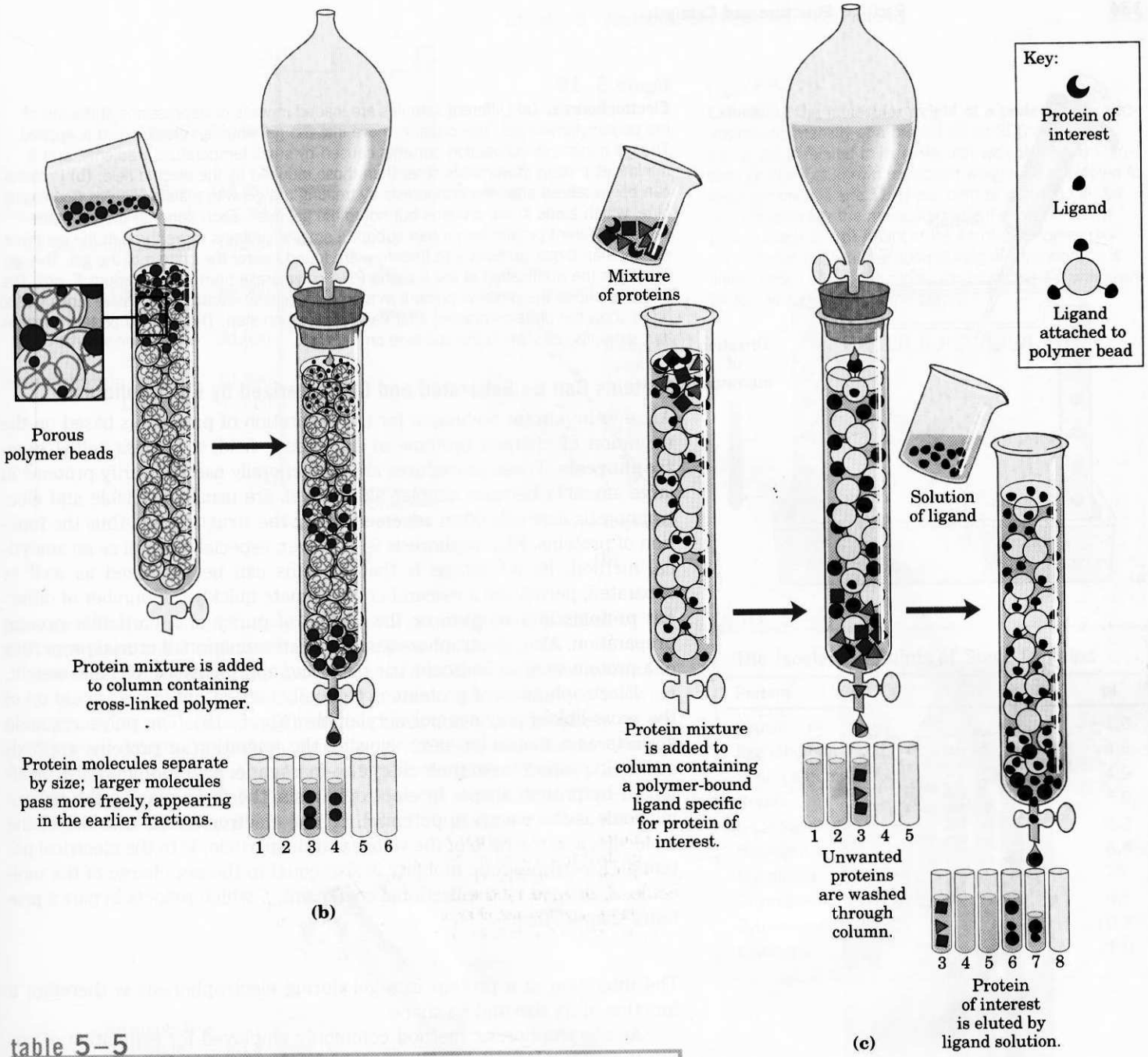
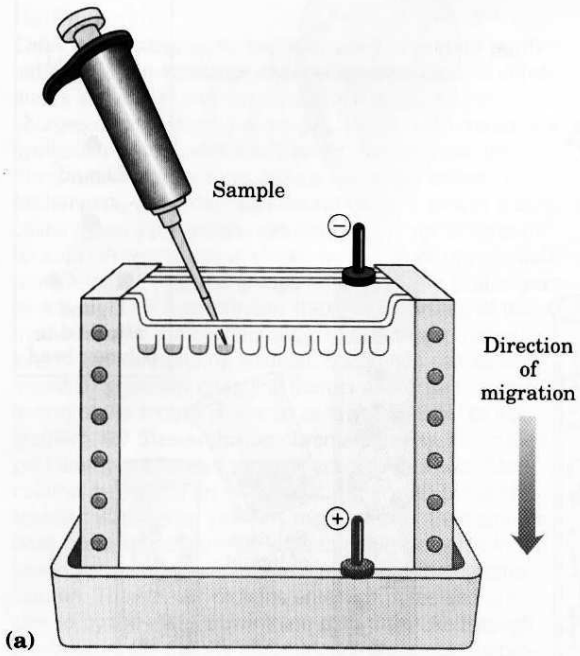


table 5–5

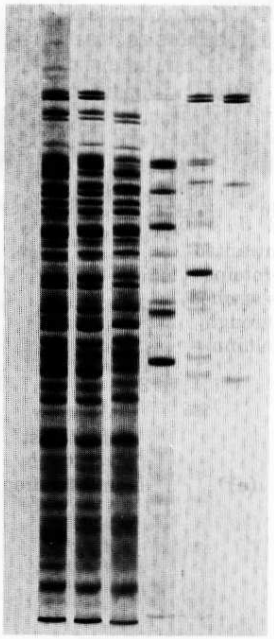
A Purification Table for a Hypothetical Enzyme*

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

*All data represent the status of the sample *after* the designated procedure has been carried out. Activity and specific activity are defined on page 137.



(a)



(b)

figure 5–19

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, and it minimizes 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 the bacterium *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.

Proteins Can Be Separated and Characterized by Electrophoresis

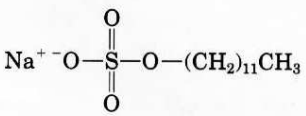
Another important technique for the separation of proteins is based on the migration of charged proteins in an electric field, a process called **electrophoresis.** These procedures are not generally used to purify proteins in large amounts because simpler alternatives are usually available and electrophoretic 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. 5–19). 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, 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).** SDS binds to most proteins (probably by hydrophobic interactions; see Chapter 4) 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. 5–19b). Thus one can monitor the progress of a protein purification procedure, because the number of protein bands visible on the gel should decrease 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. 5–20). If the protein has two



Sodium dodecyl sulfate (SDS)

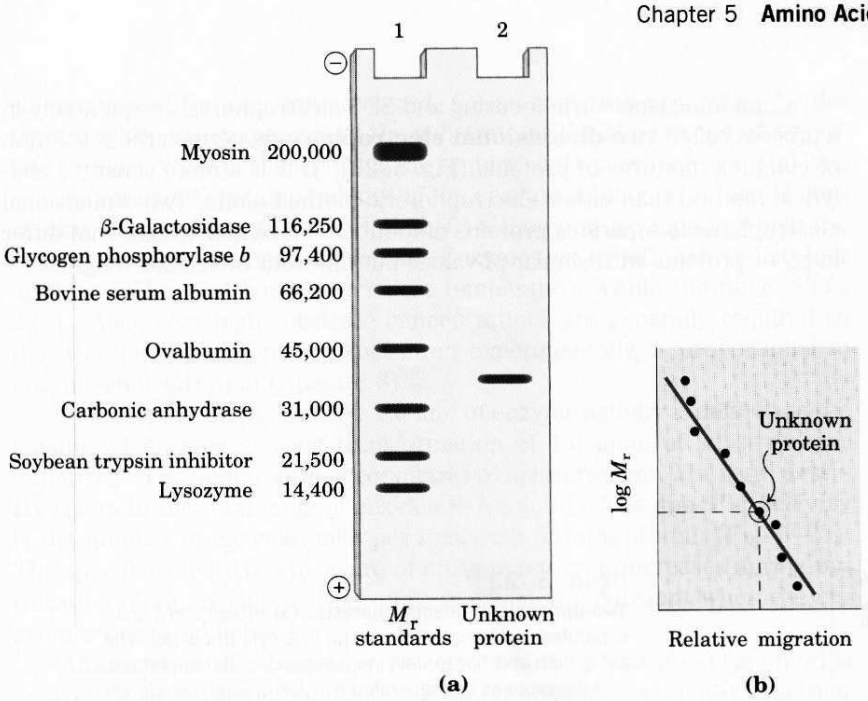


figure 5–20
Estimating the molecular weight of a protein. The electrophoretic mobility of a protein on an SDS polyacrylamide gel is related to its molecular weight, M_r. **(a)** Standard proteins of known molecular weight are subjected to electrophoresis (lane 1). These marker proteins can be used to 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.

or more different subunits, the subunits will generally be separated by the SDS treatment and a separate band will appear for each.

Isoelectric focusing is a procedure used to determine the isoelectric point (pI) of a protein (Fig. 5–21). A pH gradient is established by allowing a mixture of low molecular weight organic acids and bases (ampholytes; see p. 123) 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 5–6). Proteins with different isoelectric points are thus distributed differently throughout the gel.

table 5–6

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

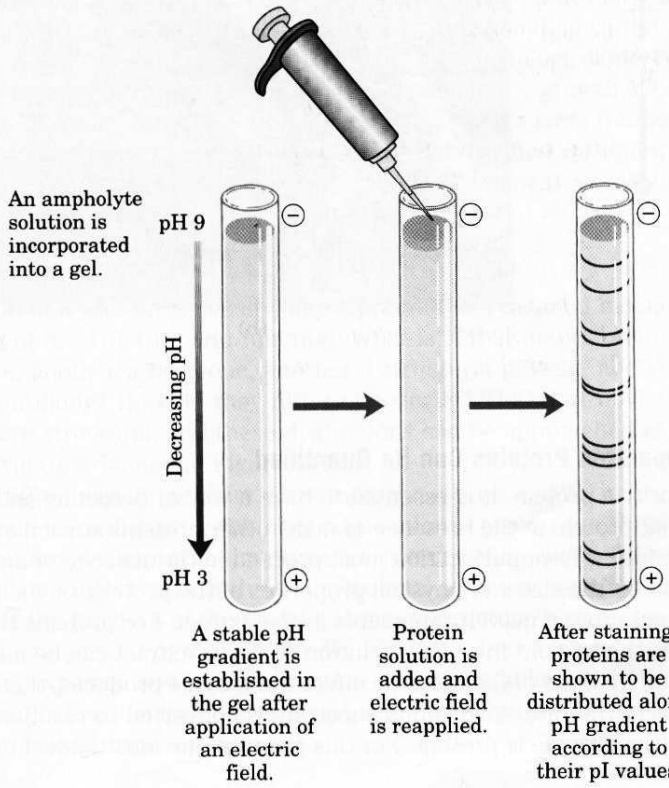
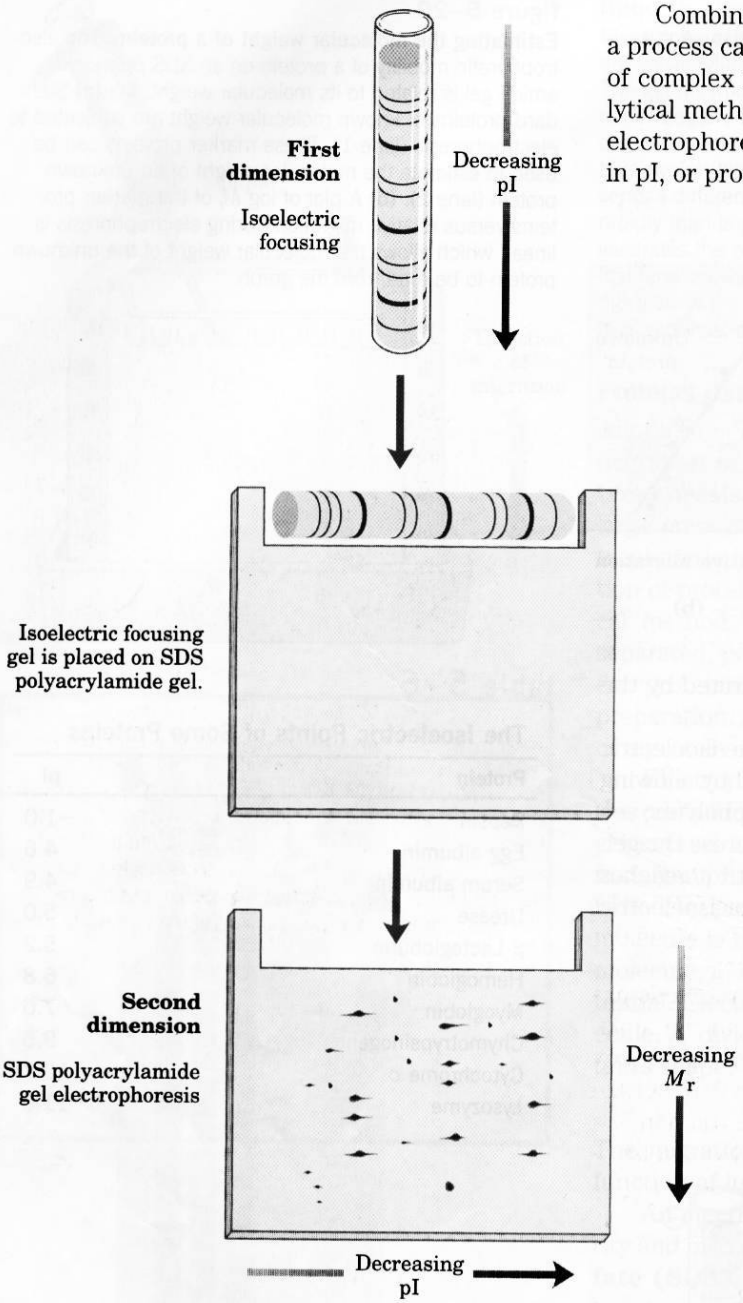


figure 5–21
Isoelectric focusing. This technique separates proteins according to their isoelectric points. A stable pH gradient is established in the gel by the addition of appropriate ampholytes. A protein mixture is placed in a well on the gel. With an applied electric field, proteins enter the gel and migrate until each reaches a pH equivalent to its pI. Remember that when pH = pI, the net charge of a protein is zero.

Combining isoelectric focusing and SDS electrophoresis sequentially in a process called **two-dimensional electrophoresis** permits the resolution of complex mixtures of proteins (Fig. 5–22). 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.



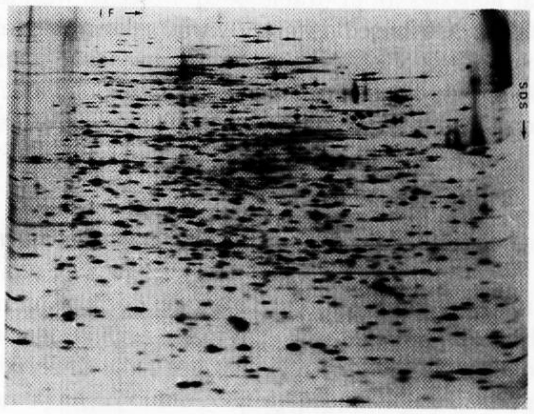


figure 5–22
Two-dimensional electrophoresis. (a) Proteins are first separated by isoelectric focusing in a cylindrical gel. The gel is then laid horizontally on a second, slab-shaped gel, and the proteins are separated by SDS polyacrylamide gel electrophoresis. Horizontal separation reflects differences 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.

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 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 over-

all 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 required so that the initial reaction rate, measured experimentally, is proportional to enzyme concentration (Chapter 8).

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. 5–23). 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 5–5).

After each purification step, the activity of the preparation (in units) is assayed, the total amount of protein is determined independently, and their ratio 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 5–5. 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.

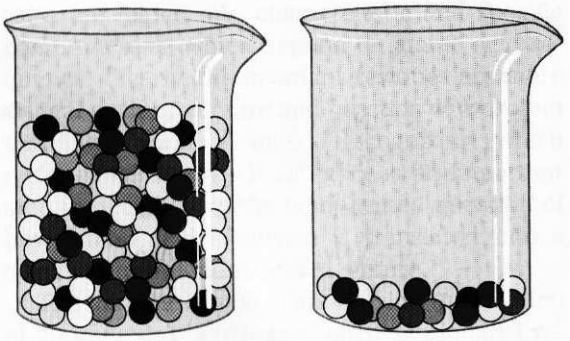
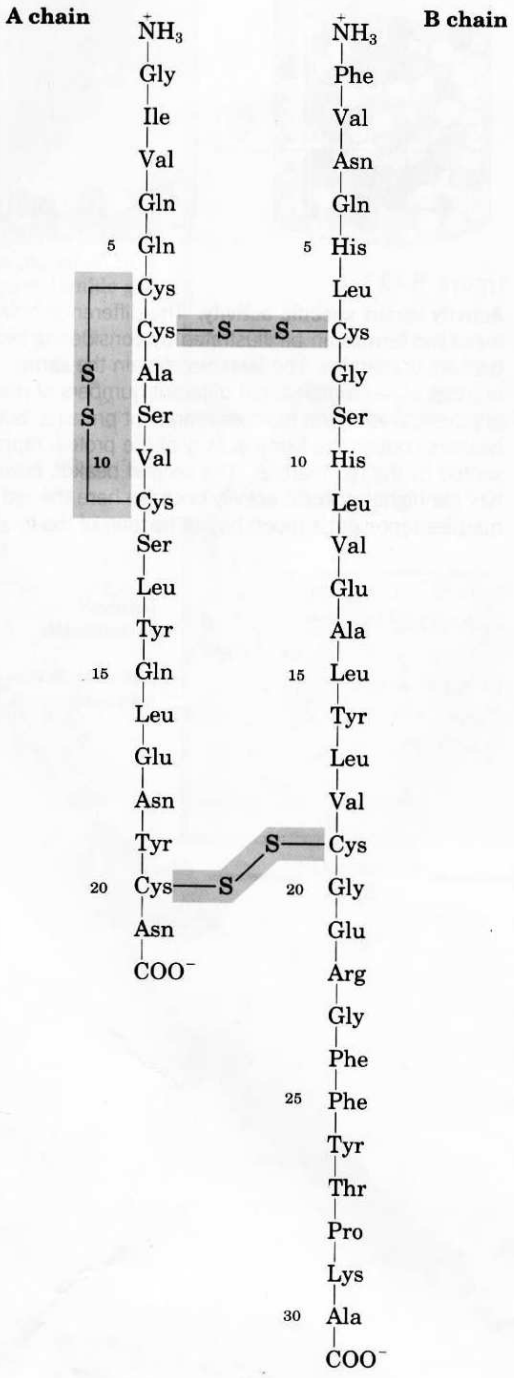


figure 5–23
Activity versus specific activity. The difference between these two terms can be illustrated by considering two beakers of marbles. The beakers contain the same number of red marbles, but different numbers of marbles of other colors. If the marbles represent proteins, both beakers contain 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.

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 5–16.

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 6, 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 now becomes the focus of the remainder of the chapter. We first consider empirical clues that amino acid sequence and protein function are closely linked, then describe how amino acid sequence is determined, and finally outline the many uses to which this information can be put.



The Function of a Protein Depends on Its Amino Acid Sequence

The bacterium *E. coli* produces more than 3,000 different proteins; a human being produces 50,000 to 100,000. 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 expectation 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 (Box 5–2). 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.

Proteins often contain crucial regions within their amino acid sequence that are essential to their biological functions. The amino acid sequence in other regions might vary considerably without affecting these functions. The fraction of the 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 Numerous 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 10). 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. 5–24), 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

figure 5–24

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. Such identities between similar proteins of different species are discussed in Box 5–2.

box 5–2 Protein Homology among Species

Homologous proteins are proteins that are evolutionarily related. They usually perform the same function in different species; an example is **cytochrome *c*,** an iron-containing mitochondrial protein that transfers electrons during biological oxidations in eukaryotic cells. Homologous proteins from different species may have polypeptide chains that are identical or nearly identical in length. Many positions in the amino acid sequence are occupied by the same residue in all species and are thus called **invariant residues.** Other positions show considerable variation in the amino acid residue from one species to another; these are called **variable residues.**

The functional significance of sequence homology is well illustrated by cytochrome *c* (M_r ~13,000), which has about 100 amino acid residues in most species. The amino acid sequences of cytochrome *c* molecules from many different species have been determined, and 27 positions in the chain are invariant in all species tested (Fig. 1), suggesting that they are the most important residues specifying the biological activity of this protein. The residues in other positions exhibit some interspecies variation. There are clear gradations in the number of differences observed in the variable residues. In some positions, most substitutions involve similar amino acid residues (for example, positively charged Arg might replace positively charged Lys); these are called **conservative substitutions.** At other positions the substitutions are less restricted (nonconservative). As we will show in the next chapter, the polypeptide chains of proteins are folded into characteristic and specific conformations, which depend on amino acid sequence. Clearly, the invariant residues are more critical to the structure and function of a protein than the variable ones. Recognizing which residues fall into each category is an important step in deciphering the complicated question of how amino acid sequence is translated into a specific three-dimensional structure.

The variable amino acids provide information of another sort. Phylogenetic (evolutionary) relationships based on taxonomic methods have been tested and experimentally confirmed through biochemistry. The examination of sequences of cytochrome *c* and other homologous proteins has led to an important conclusion: the number of residues that differ in homologous proteins from any two species is in proportion to the phylogenetic difference between those species. For example, 48 amino acid residues differ in the cytochrome *c* molecules of the horse and of yeast, which are very widely separated species, whereas only two residues differ in the cytochrome *c* molecules of the much more closely related duck and chicken. In fact, cytochrome *c* has identical amino acid sequences in the chicken and the turkey, and in the pig, cow,

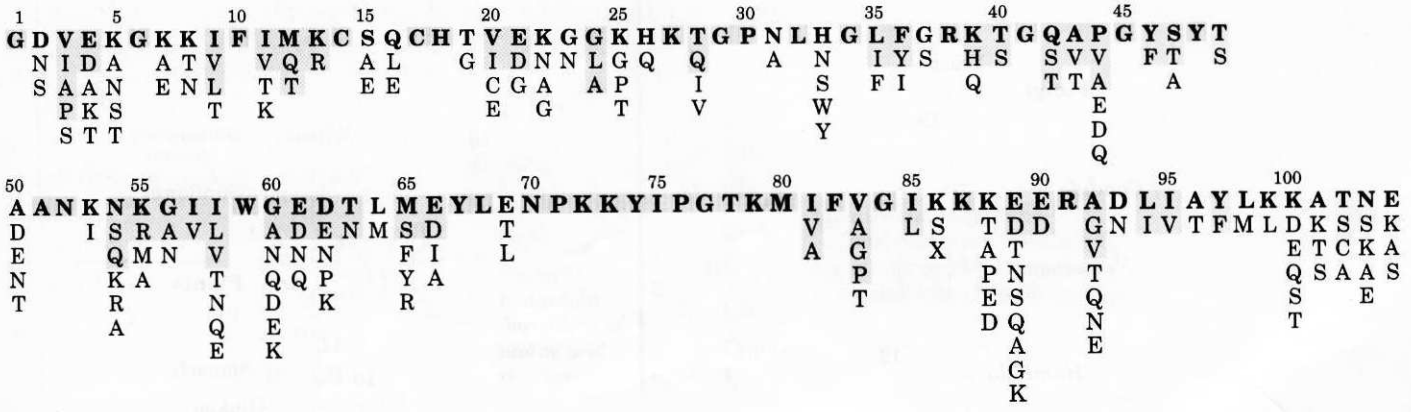


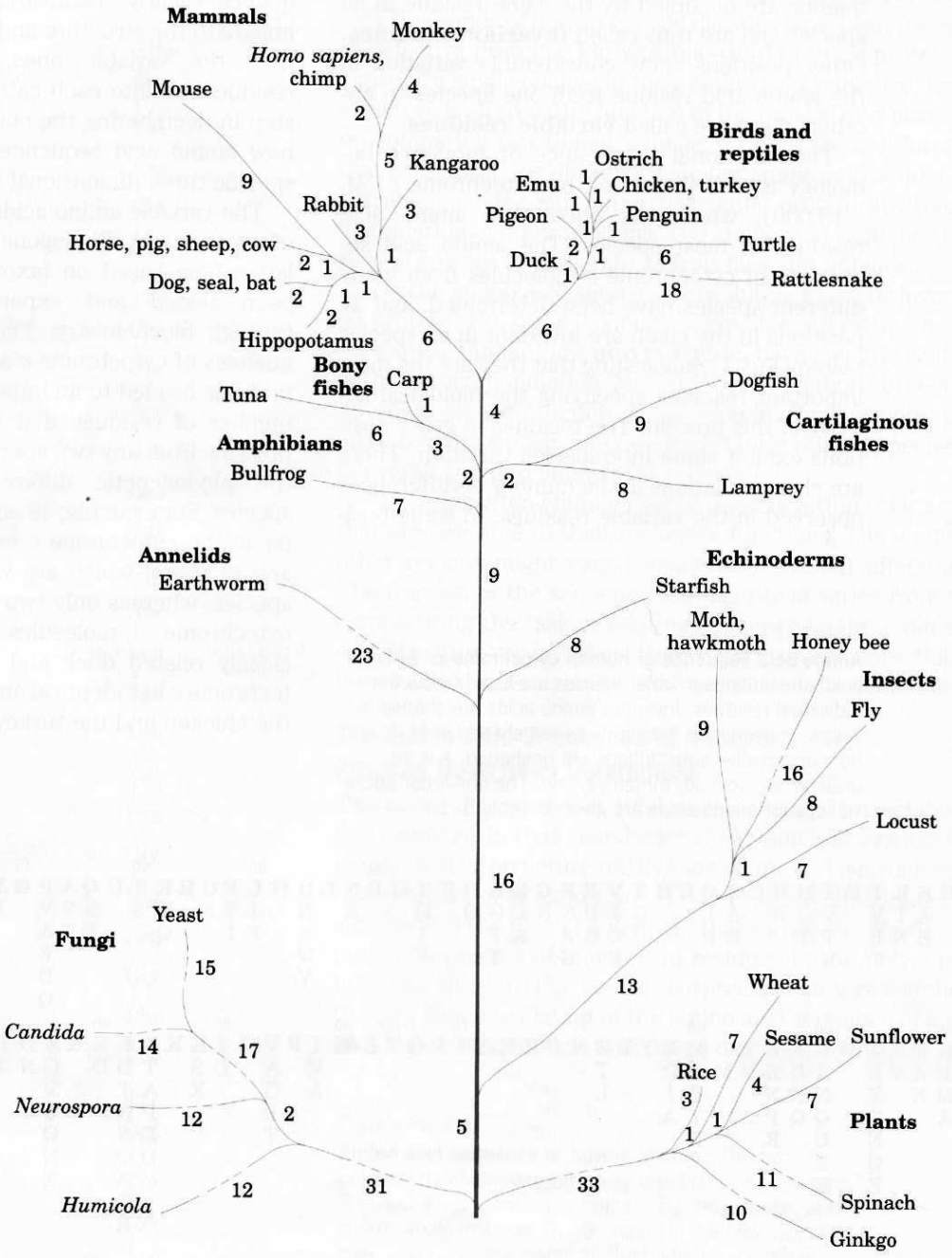
figure 1
Amino acid sequence of human cytochrome *c*. Amino acid substitutions in other species are listed below the individual residues. Invariant amino acids are shaded in yellow, conservative substitutions are shaded in blue, and nonconservative substitutions are unshaded. X is an unusual amino acid, trimethyllysine. The one-letter abbreviations for amino acids are given in Table 5–1.

Continued on next page

and sheep. Information on the number of residue differences between homologous proteins of different species allows the construction of evolutionary trees that show the origin and sequence of appearance of different species during the course of evolution (Fig. 2). The relationships established by anatomic and biochemical taxonomy are in close agreement.

figure 2

Main branches of the eukaryotic evolutionary tree constructed from the number of amino acid differences between cytochrome *c* molecules of various species. The numbers represent the number of residues by which the cytochrome *c* of a given line of organism differs from its ancestor. Branch points reflect a common ancestor.



sequence in DNA and the amino acid sequence in proteins were somehow related. Barely a decade after these discoveries, the role of DNA nucleotide sequence in determining the amino acid sequence of protein molecules had been revealed (Chapter 27).

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.

Short Polypeptides Are Sequenced Using Automated Procedures

Various procedures are used to analyze protein primary structure. One is to hydrolyze the protein and determine its amino acid composition (Fig. 5–25a). This information is often valuable in interpreting the results of other procedures. Because amino acid composition differs from one protein to the next, it can serve as a kind of fingerprint. It can be used, for example, to help determine whether proteins isolated by different laboratories are the same or different. Hydrolysis alone, however, cannot be used to determine the *sequence* of amino acids in a protein.

A procedure that is often used in conjunction with hydrolysis is to label and identify the amino-terminal amino acid residue (Fig. 5–25b). For this purpose Sanger developed the reagent 1-fluoro-2,4-dinitrobenzene (FDNB).

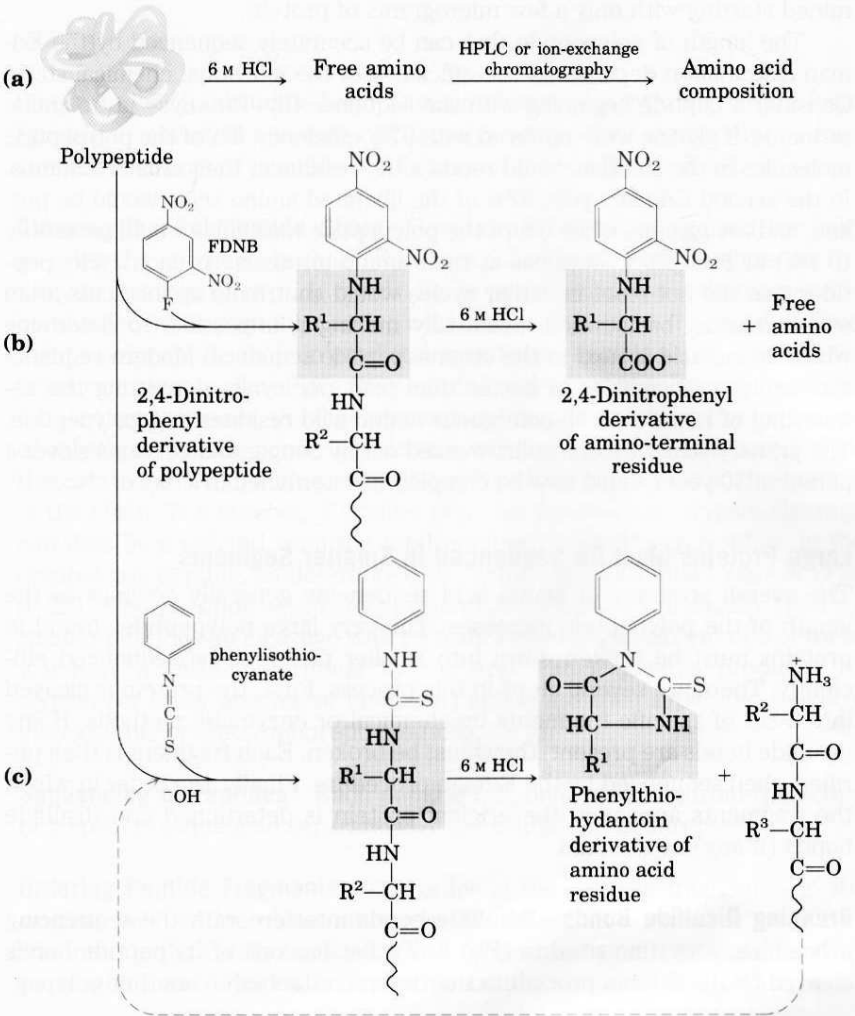
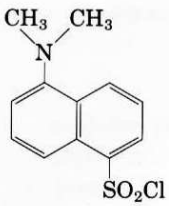
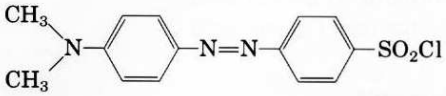


figure 5–25

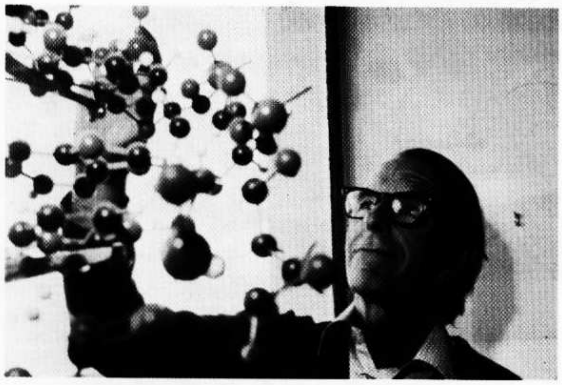
Steps in sequencing a polypeptide. (a) Determination of amino acid composition and **(b)** identification of the amino-terminal residue are often the first steps in sequencing a polypeptide. Sanger's method for identifying the amino-terminal residue is shown here. The Edman degradation procedure **(c)** reveals the entire sequence of a peptide. For shorter peptides, this method alone readily yields the entire sequence, and steps **(a)** and **(b)** are often omitted. The latter procedures are useful in the case of larger polypeptides, which are often fragmented into smaller peptides for sequencing (see Fig. 5–27).



Dansyl chloride



Dabsyl chloride



Frederick Sanger

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. 5–24) were subjected to this procedure.

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. 5–25c). The peptide is reacted with phenylisothiocyanate, and the amino-terminal residue is ultimately removed as a phenylthiohydantoin derivative. 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,** which 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 then 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. 5–7) that has one of its peptide bonds cleaved by the Edman procedure may remain attached to another polypep-

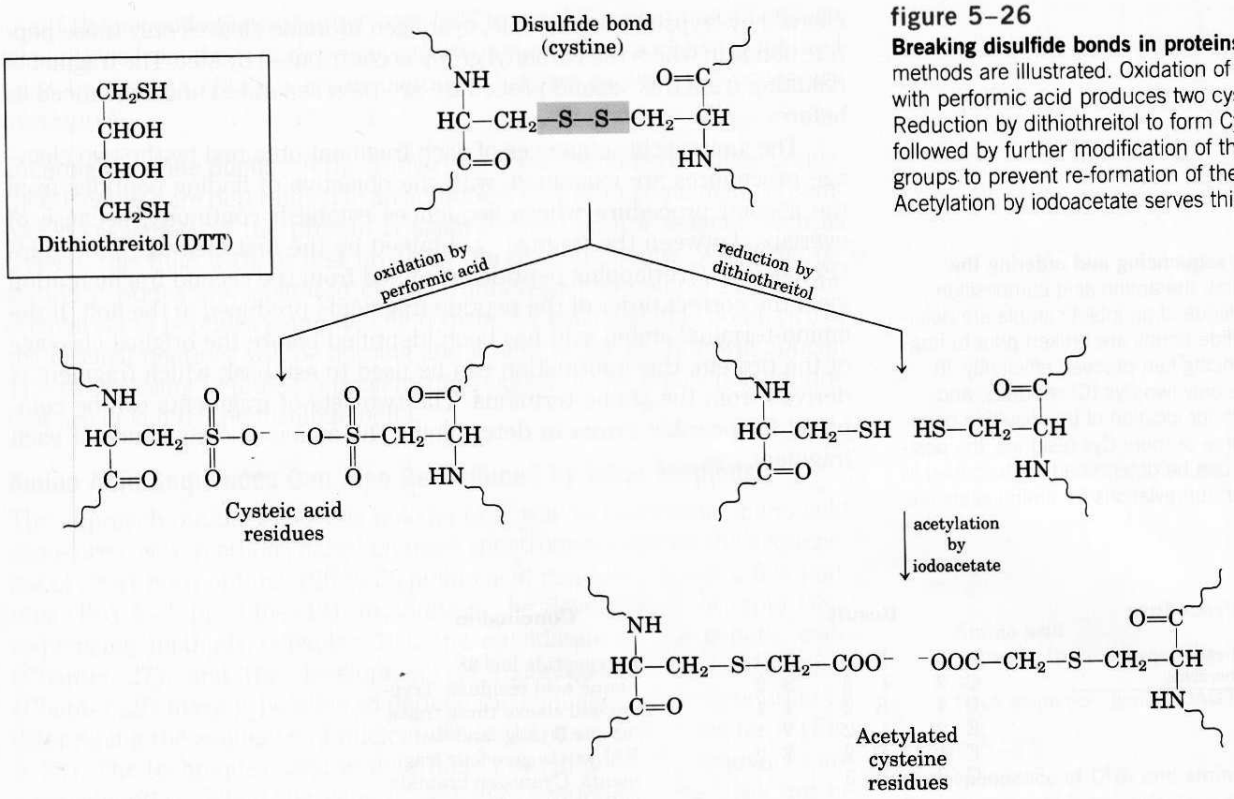


figure 5–26
Breaking disulfide bonds in proteins. Two common methods are illustrated. Oxidation of a cystine residue with performic acid produces two cysteic acid residues. Reduction by dithiothreitol to form Cys residues must be followed by further modification of the reactive —SH groups to prevent re-formation of the disulfide bond. Acetylation by iodoacetate serves this purpose.

tide 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 5–26.

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 5–7) and thus fragment a polypeptide chain in predictable and reproducible ways. 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 from the total number of Lys or Arg residues in the original polypeptide, as determined by hydrolysis of an intact sample (Fig. 5–27). 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

table 5–7

The Specificity of Some Common Methods for Fragmenting Polypeptide Chains

Treatment*	Cleavage points†
Trypsin	Lys, Arg (C)
Submaxillarus protease	Arg (C)
Chymotrypsin	Phe, Trp, Tyr (C)
Staphylococcus aureus V8 protease	Asp, Glu (C)
Asp-*N*-protease	Asp, Glu (N)
Pepsin	Phe, Trp, Tyr (N)
Endoproteinase Lys C	Lys (C)
Cyanogen bromide	Met (C)

*All 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.

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. 5–27). 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.

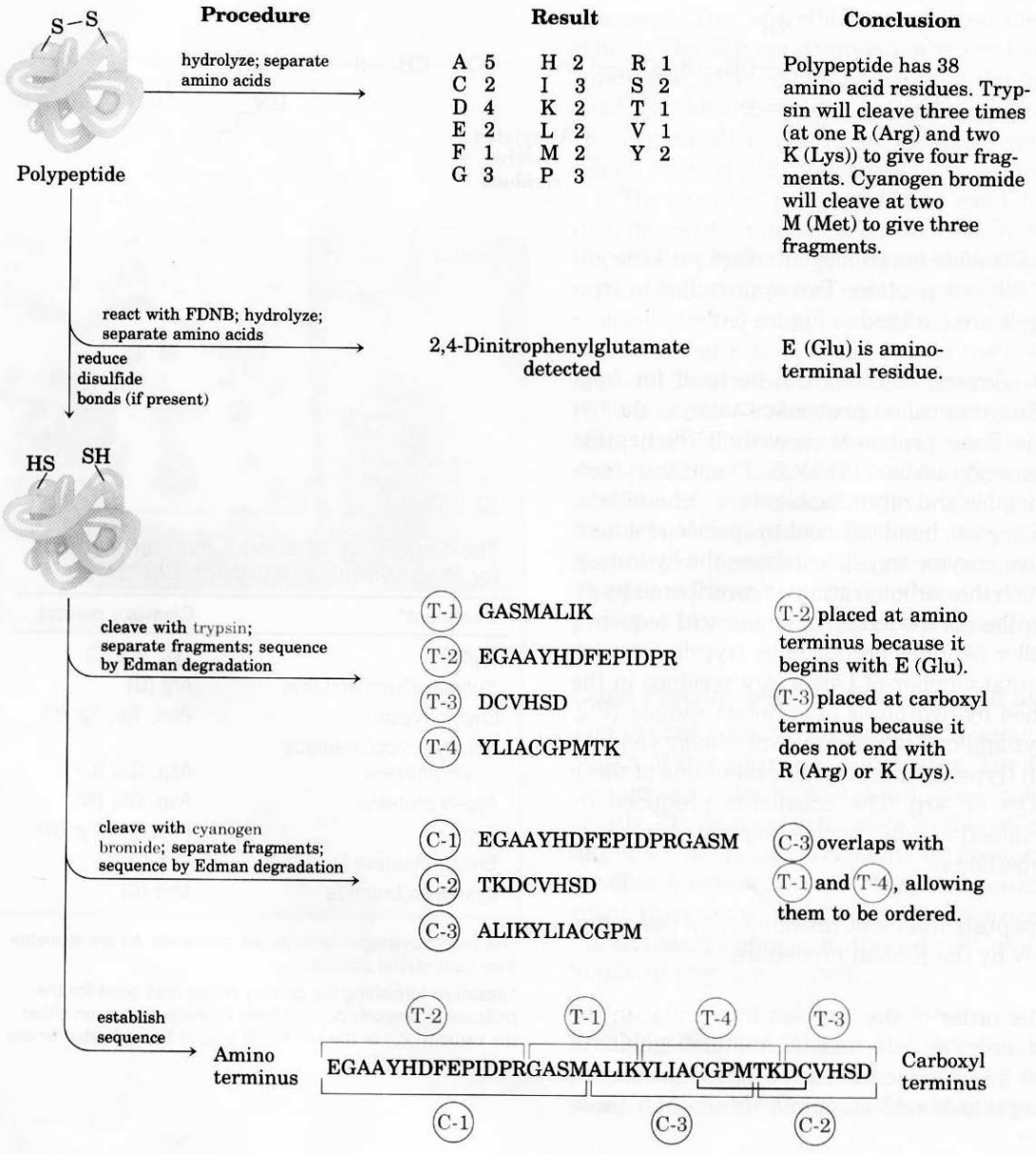


figure 5–27
Cleaving proteins and sequencing and ordering the peptide fragments. First, the amino acid composition and amino-terminal residue of an intact sample are determined. Then any disulfide bonds are broken prior to fragmenting so that sequencing can proceed efficiently. In this example, there 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 abbreviations for amino acids are given in Table 5–1.)

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. When 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 5–3, pp. 146–149). In addition, the development of rapid DNA sequencing methods (Chapter 10), the elucidation of the genetic code (Chapter 27), and the development of techniques for isolating genes (Chapter 29) make it possible to deduce the sequence of a polypeptide by determining the sequence of nucleotides in the gene that codes for it (Fig. 5–28). 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 (e.g., the location of disulfide bonds) 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 29).

The array of methods now available to analyze both proteins and nucleic acids is ushering in a new discipline 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 1–1). Genes are being discovered by the thousands, 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.** Analysis of a cell's proteome is an increasingly important and informative adjunct to the completion of its genomic sequence. Proteins from a cell are separated and displayed by two-dimensional gel electrophoresis (Fig. 5–22). Individual protein spots can be extracted from such a gel. Small peptides derived from the proteins are sequenced by mass spectrometry (Box 5–3), and these sequences are compared with the genomic sequence to identify the protein. Often, knowledge of the sequence of a segment of six to eight amino acid residues is enough to pinpoint the gene encoding the entire protein. Inevitably, some of these proteins are already known and have well-studied functions; others are more mysterious. Once most of the proteins are matched to a gene, changes in a cell's protein complement brought on by the environment, nutritional changes, stress, or disease can be examined. This work can provide clues to the role of proteins whose functions are as yet unknown. Eventually, such studies will complement 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.

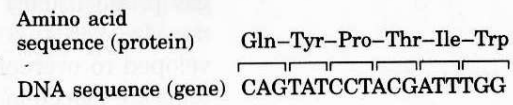


figure 5–28
Correspondence of DNA and amino acid sequences. Each amino acid is encoded by a specific sequence of three nucleotides in DNA. The genetic code is described in detail in Chapter 27.

box 5–3 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 bring a macromolecule into 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 proteome 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,

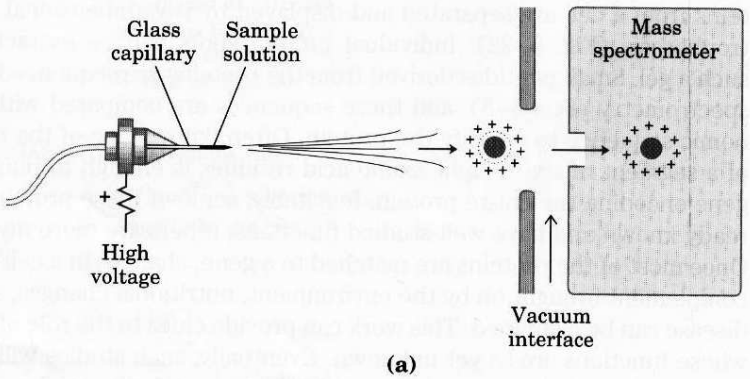


figure 1
Electrospray mass spectrometry of a protein. (a) A protein solution is dispersed into highly charged droplets by passage through a needle under the influence of a high-voltage electric field. The droplets evaporate, and the ions (with added protons in this case) enter the mass spectrometer for 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.

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 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,

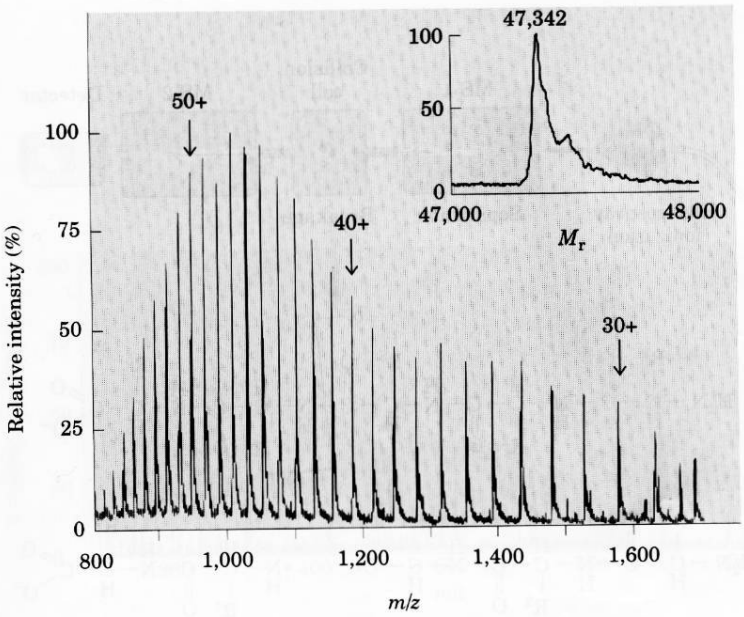
$$(m/z)_1 = \frac{M + (n_2 + 1)X}{n_2 + 1}$$

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 will usually provide 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).



(b)

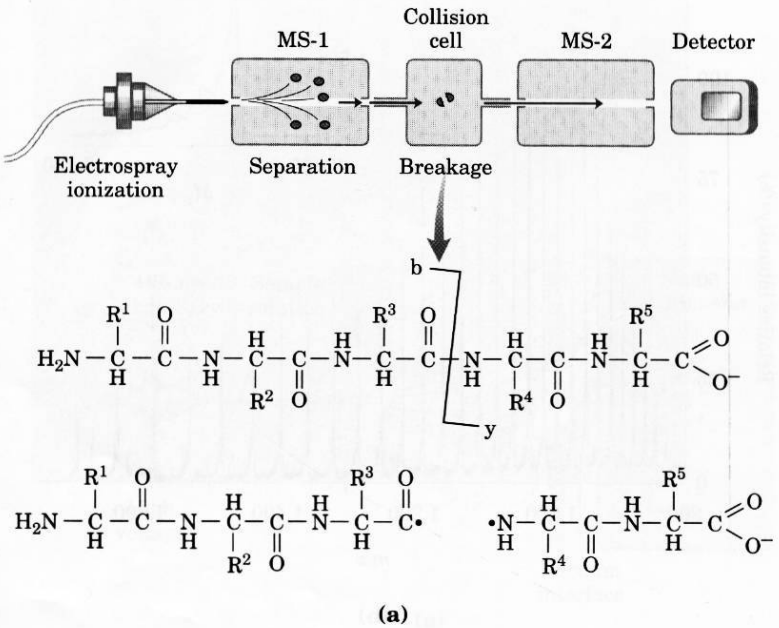
Continued on next page

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 reduce it by hydrolytic cleavage 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 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 derived from the same side of the bond breakage, either the carboxyl-terminal 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

figure 2

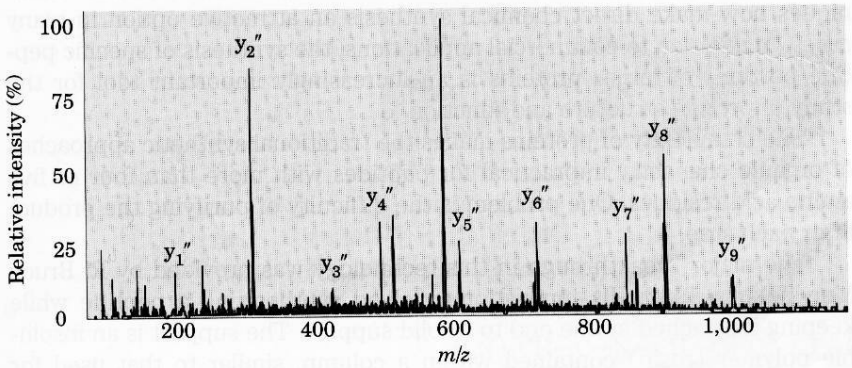
Obtaining protein sequence information with tandem mass spectrometry. (a) After proteolytic hydrolysis, a protein solution is injected into a mass spectrometer (MS-1). The different peptides are sorted so that only one type is selected for further analysis. The selected peptide is further fragmented 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.



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 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 assigned 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 proteome 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.



(b)

Amino Acid Sequences Provide Important Biochemical Information

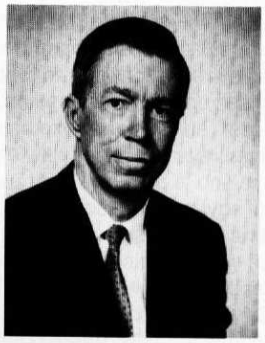
Knowledge of the sequence of amino acids in a protein can offer insights into its three-dimensional structure and its function, cellular location, and evolution. Most of these insights are derived by searching for similarities with other known sequences. Thousands of sequences are known and available in databases accessible through the Internet. The 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 6 as "domains") occur in many functionally unrelated proteins. These domains often fold up 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 often 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, while 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 if they occur in the sequence of a newly characterized protein.

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 29); 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.



R. Bruce Merrifield

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. 5–29). At each successive step in the cycle, protective chemical groups block unwanted reactions.

figure 5–29

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).

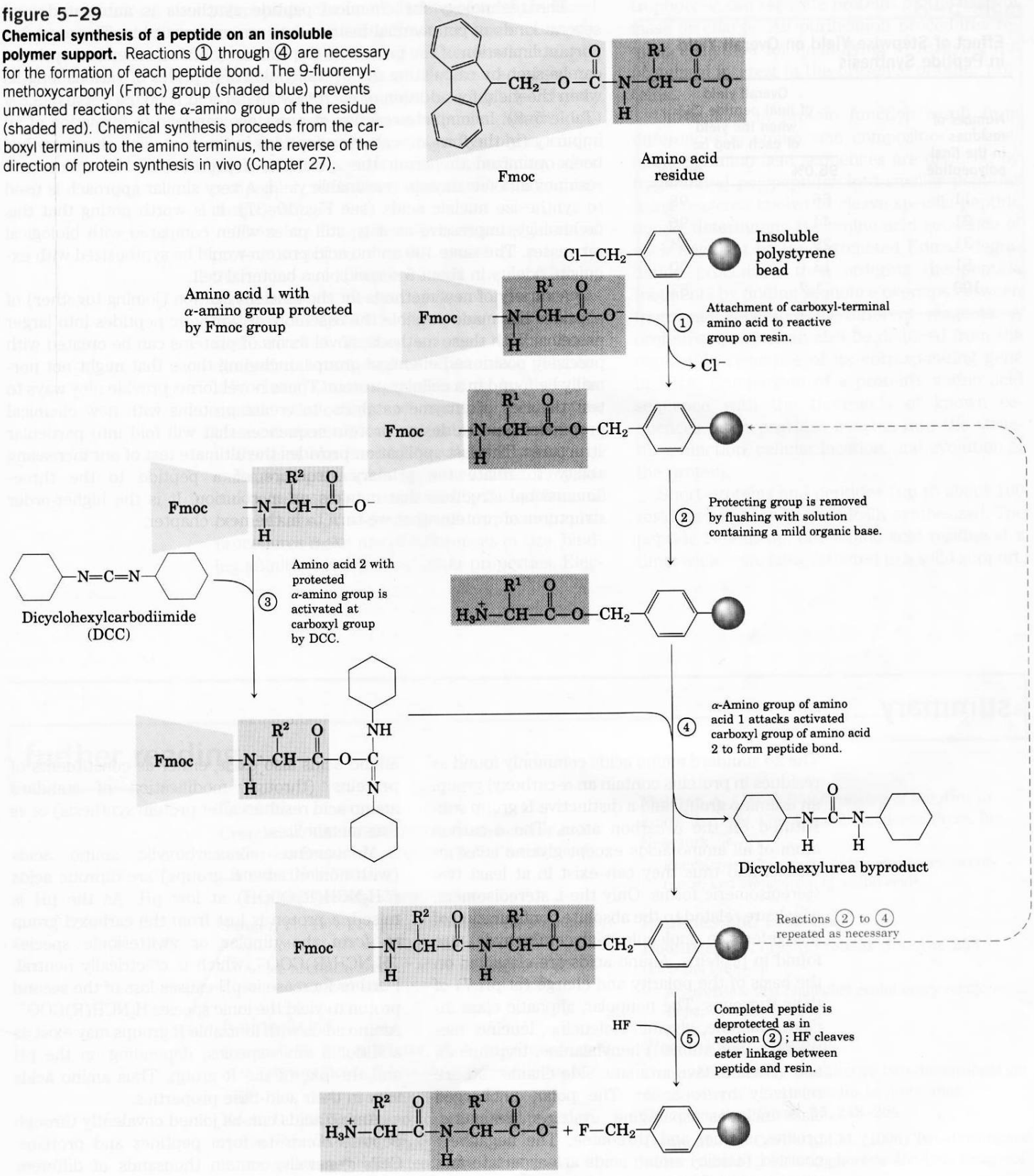


table 5–8

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

The technology for chemical peptide synthesis is automated, and several kinds of commercial instruments are now available. 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 5–8). 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. 10–37). It is worth noting that this technology, impressive as it is, still pales when compared with biological processes. The same 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. It is the higher-order structures of proteins that we turn to in the next chapter.

summary

The 20 standard 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 they can exist in at least two stereoisomeric forms. Only the L stereoisomers, which are related to the absolute configuration of the reference molecule L-glyceraldehyde, are found in proteins. Amino acids are classified on the basis of the polarity and charge (at pH 7) of their R groups. The nonpolar, aliphatic class includes alanine, glycine, isoleucine, leucine, methionine, and valine. Phenylalanine, tryptophan, and tyrosine have aromatic side chains and are relatively hydrophobic. The polar, uncharged class includes asparagine, cysteine, glutamine, proline, serine, and threonine. The negatively charged (acidic) amino acids are aspartate and glutamate; the positively charged (basic) ones are arginine, histidine, and lysine. Nonstandard amino acids also exist, either as constituents of proteins (through modification of standard amino acid residues after protein synthesis) or as free metabolites.

Monoamino monocarboxylic amino acids (with nonionizable R groups) are diprotic acids ($^+H_3NCH(R)COOH$) at low pH. As the pH is raised, a proton is lost from the carboxyl group to form the dipolar or zwitterionic species $^+H_3NCH(R)COO^-$, which is electrically neutral. Further increase in pH causes loss of the second proton to yield the ionic species $H_2NCH(R)COO^-$. Amino acids with ionizable R groups may exist as additional ionic species, depending on the pH and the pK_a of the R group. Thus amino acids vary in their acid-base properties.

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

There are four generally recognized levels of protein structure. Primary structure refers to the amino acid sequence and the location of disulfide bonds. Secondary structure is the spatial relationship of adjacent amino acids in localized stretches. Tertiary structure is the three-dimensional conformation of an entire polypeptide chain. Quaternary structure involves the spatial relationship of multiple polypeptide chains (subunits) that are stably associated.

Proteins are purified by taking advantage of various properties in which they differ. Proteins can be selectively precipitated by the addition of certain salts. A wide range of chromatographic procedures make use of differences in size, binding affinities, charge, and other properties. Electrophoresis can separate proteins on the basis of mass or charge. All purification procedures require a method for quantifying or assaying the protein of interest in the presence of other proteins.

Differences in protein function result from differences in amino acid composition and sequence. 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. Comparison of a protein's amino acid sequence with the thousands of known sequences often provides insights into the structure, function, cellular location, and evolution of the protein.

Short proteins and peptides (up to about 100 residues long) can be chemically synthesized. The peptide is built up, one amino acid residue at a time, while remaining tethered to a solid support.

further reading

General

Creighton, T.E. (1993) *Proteins: Structures and Molecular Properties,* 2nd edn, W.H. Freeman and Company, New York.

Very useful general source.

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

A nice historical account of the development of sequencing methods.

Amino Acids

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

Doolittle, R.F. (1985) Proteins. *Sci. Am.* **253** (October), 88–99.

An overview that highlights evolutionary relationships.

Working with Proteins

Dunn, M.J. (1997) Quantitative two-dimensional gel electrophoresis: from proteins to proteomes. *Biochem. Soc. Trans.* **25,** 248–254.

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

Andersen, J.S., Svensson, B., & Roepstorff, P. (1996) Electrospray ionization and matrix assisted laser desorption/ionization mass spectrometry: powerful analytical tools in recombinant protein chemistry. *Nat. Biotechnol.* **14,** 449–457.

A summary emphasizing applications.

Bork, P. & Koonin, E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nat. Genet.* **18,** 313–318.

A good description of the technology and the roadblocks still limiting its use.

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 methods.

Gibney, B.R., Rabanal, F., & Dutton, P.L. (1997) Synthesis of novel proteins. *Curr. Opin. Chem. Biol.* **1,** 537–542.

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 of what we will do with 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 for beginners.

Wallace, C.J. (1995) Peptide ligation and semisynthesis. *Curr. Opin. Biotechnol.* **6,** 403–410.

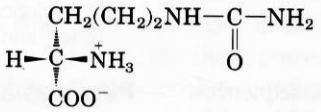
Good summary of methods available for peptide ligation. Includes some case studies

Wilken, J. & Kent, S.B. (1998) Chemical protein synthesis. *Curr. Opin. Biotechnol.* **9,** 412–426.

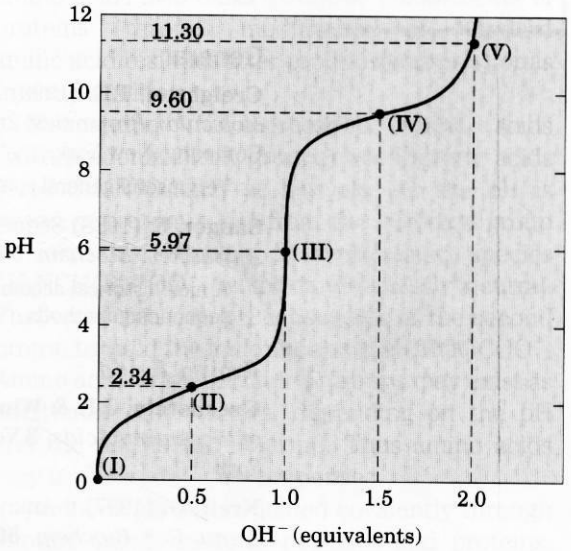
A good overview of chemical synthesis, focusing on peptide ligation methods and applications.

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.

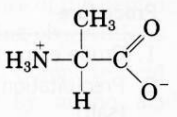
(l) Glycine is present predominantly as a 50:50 mixture of $^{+}H_3N{-}CH_2{-}COOH$ and $^{+}H_3N{-}CH_2{-}COO^{-}$.

(m) This is the isoelectric point.

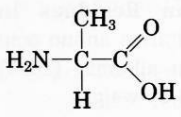
(n) This is the end of the titration.

(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. 5–18) 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

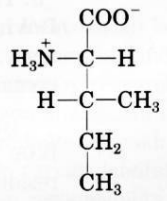
(b) Arg and Met

(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 5–1.)

(b) Estimate the pI for this peptide.

11. Isoelectric Point of Pepsin Pepsin is the name given to several digestive enzymes that are 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 polypeptides below is more soluble at the indicated pH?

(a) $(Gly)_{20}$ or $(Glu)_{20}$ at pH 7.0

(b) $(Lys\text{-}Ala)_3$ or $(Phe\text{-}Met)_3$ at pH 7.0

(c) $(Ala\text{-}Ser\text{-}Gly)_5$ or $(Asn\text{-}Ser\text{-}His)_5$ at pH 6.0

(d) $(Ala\text{-}Asp\text{-}Gly)_5$ or $(Asn\text{-}Ser\text{-}His)_5$ 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?