

Molecular ribbon model of a protein “needle” used by pathogenic bacteria to inject proteins into human cells to initiate infection. Many disease-causing bacteria, including *Salmonella typhimurium* (food poisoning) and *Yersinia pestis* (bubonic plague), use a syringe-like protein complex called a type III secretion system to inject proteins into their mammalian target cells. The structure of the needle portion of the syringe used by *Salmonella typhimurium*, determined using a combination of nuclear magnetic resonance (NMR), electron microscopy, and computational methods, is a long tube with many  $\alpha$  helices (illustrated as coiled ribbons) forming the walls of the needle. [Data from A. Loquet et al., 2012, *Nature* **486**:276, PDB ID 2lpz.]

Proteins, which are polymers of amino acids, come in many sizes and shapes. Their three-dimensional diversity principally reflects variations in their lengths and amino acid sequences. In general, the linear, unbranched polymer of amino acids composing any protein will fold into only one or a few closely related three-dimensional shapes—called **conformations**. The conformation of a protein, together with the distinctive chemical properties of its amino acid side chains, determines its function. In some cases, the conformation, and thus the function, of a protein can change when that protein noncovalently or covalently associates with other molecules. Because of their many different shapes and chemical properties, proteins can perform a dazzling array of distinct functions inside and outside cells that either are essential for life or provide a

# Protein Structure and Function

selective evolutionary advantage to the cell or organism that contains them. It is, therefore, not surprising that characterizing the structures and activities of proteins is a fundamental prerequisite for understanding how cells work. Much of this textbook is devoted to examining how proteins act together to allow cells to live and function properly.

Although their structures are diverse, most proteins can be grouped into one of a few broad functional classes. *Structural proteins*, for example, determine the shapes of cells and their extracellular environments and serve as guide wires or rails to direct the intracellular movement of molecules and organelles. They are usually formed by the assembly of multiple protein subunits into very large, long structures. *Scaffold proteins* bring other proteins together into ordered

## OUTLINE

### 3.1 Hierarchical Structure of Proteins

### 3.2 Protein Folding

### 3.3 Protein Binding and Enzyme Catalysis

### 3.4 Regulating Protein Function

### 3.5 Purifying, Detecting, and Characterizing Proteins

### 3.6 Proteomics

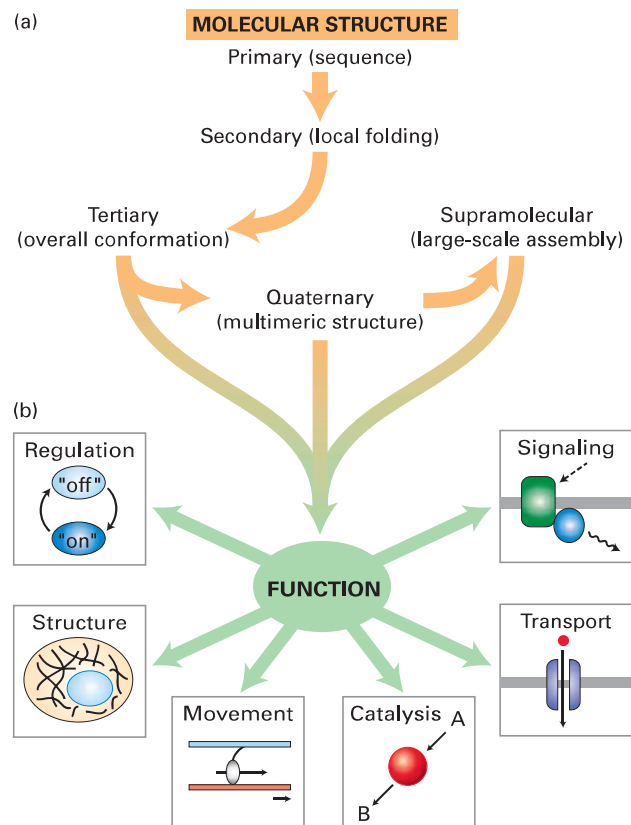
arrays to perform specific functions more efficiently than those proteins would if they were not assembled together. *Enzymes* are proteins that catalyze chemical reactions. *Membrane transport proteins* permit the flow of ions and molecules across cellular membranes. *Regulatory proteins* act as signals, sensors, and switches to control the activities of cells by altering the functions of other proteins and genes. Regulatory proteins include *signaling proteins*, such as the hormones and cell-surface receptors that transmit extracellular signals to the cell interior. *Motor proteins* are responsible for moving other proteins, organelles, cells—even whole organisms. Any one protein can be a member of more than one protein class, as is the case with some cell-surface signaling receptors that are both enzymes and regulator proteins because they transmit signals from outside to inside cells by catalyzing chemical reactions. To accomplish their diverse missions efficiently, some proteins assemble into large complexes, often called *molecular machines*.

How do proteins perform so many diverse functions? They do so by exploiting a few simple activities. Most fundamentally, proteins *bind*—to one another, to other macromolecules such as DNA, and to small molecules and ions. In many cases, such binding induces a conformational change (a change in the three-dimensional structure) in the protein and thus influences its activity. Binding is based on molecular complementarity between a protein and its binding partner, as described in Chapter 2. A second key activity is enzymatic *catalysis*. Appropriate folding of a protein will place some amino acid side chains and some carboxyl and amino groups of its backbone into positions that permit the catalysis of covalent bond rearrangements. A third activity is *folding into a channel or pore* within a membrane through which molecules and ions can flow. Although these are especially crucial protein activities, they are not the only ones. For example, fish that live in frigid waters—the Antarctic borchs and Arctic cods—have antifreeze proteins in their circulatory systems to prevent water crystallization.

A complete understanding of how proteins permit cells to live and thrive requires the identification and characterization of all the proteins used by a cell. In a sense, molecular cell biologists want to compile a complete protein “parts list” and construct a “user’s manual” that describes how these proteins work. Compiling a comprehensive inventory of proteins has become feasible in recent years with the sequencing of the entire genomes—complete sets of genes—of more and more organisms. From a computer analysis of a genome’s sequence, researchers can deduce the amino acid sequences and approximate number of the proteins it encodes (see Chapter 6). The term **proteome** was coined to refer to the entire protein complement of an organism. The human genome contains some 20,000–23,000 genes that encode proteins. However, variations in mRNA production, such as alternative splicing (see Chapter 10), and more than a hundred types of protein modifications may generate hundreds of thousands of distinct human proteins. By comparing the sequences and structures of proteins of unknown function with those of proteins of known function, scientists can often deduce much about what the unknown

proteins do. In the past, characterization of protein function by genetic, biochemical, or physiological methods often preceded the identification of particular proteins. In the modern genomic and proteomic era, a protein is usually identified before its function is determined.

In this chapter, we begin our study of how the structure of a protein gives rise to its function, a theme that recurs throughout this book (Figure 3-1). The first section examines how linear chains of amino acid building blocks are arranged in a three-dimensional structural hierarchy. The next section discusses how proteins fold into these structures. We then turn to protein function, focusing on enzymes, those proteins that catalyze chemical reactions. Various mechanisms that cells use to control the activities and life spans of proteins are covered next. The chapter concludes with a discussion

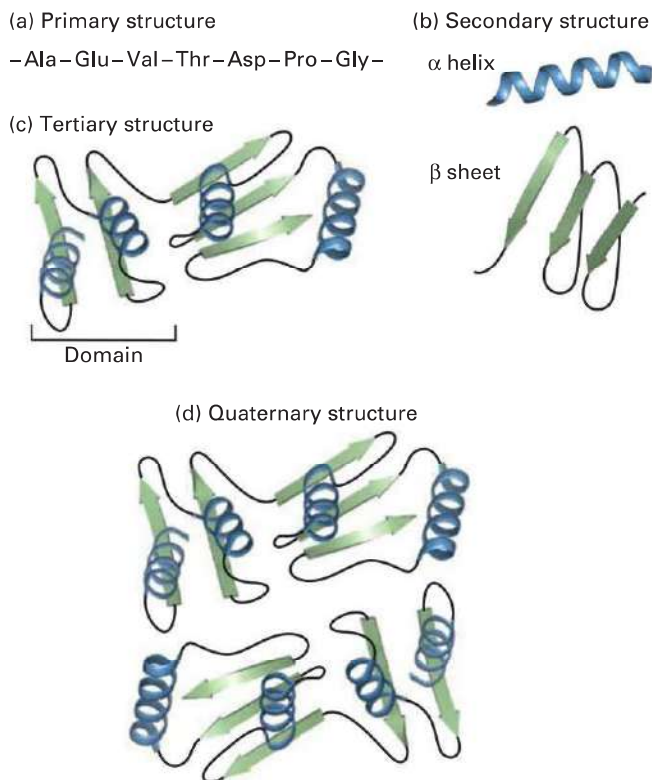


**FIGURE 3-1 Overview of protein structure and function.** (a) Proteins have a hierarchical structure. A polypeptide’s linear sequence of amino acids linked by peptide bonds (primary structure) folds into local helices or sheets (secondary structure) that pack into a complex three-dimensional shape (tertiary structure). Some individual polypeptides associate into multichain complexes (quaternary structure), which in some cases can be very large, consisting of tens to hundreds of subunits (supramolecular complexes). (b) Proteins perform numerous functions, including organizing the genome, organelles, cytoplasm, protein complexes, and membranes in three-dimensional space (structure); controlling protein activity (regulation); monitoring the environment and transmitting information (signaling); moving small molecules and ions across membranes (transport); catalyzing chemical reactions (via enzymes); and generating force for movement (via motor proteins). These functions and others arise from specific binding interactions and conformational changes in the structure of a properly folded protein.

of commonly used techniques for identifying, isolating, and characterizing proteins, and a discussion of the burgeoning field of proteomics.

### 3.1 Hierarchical Structure of Proteins

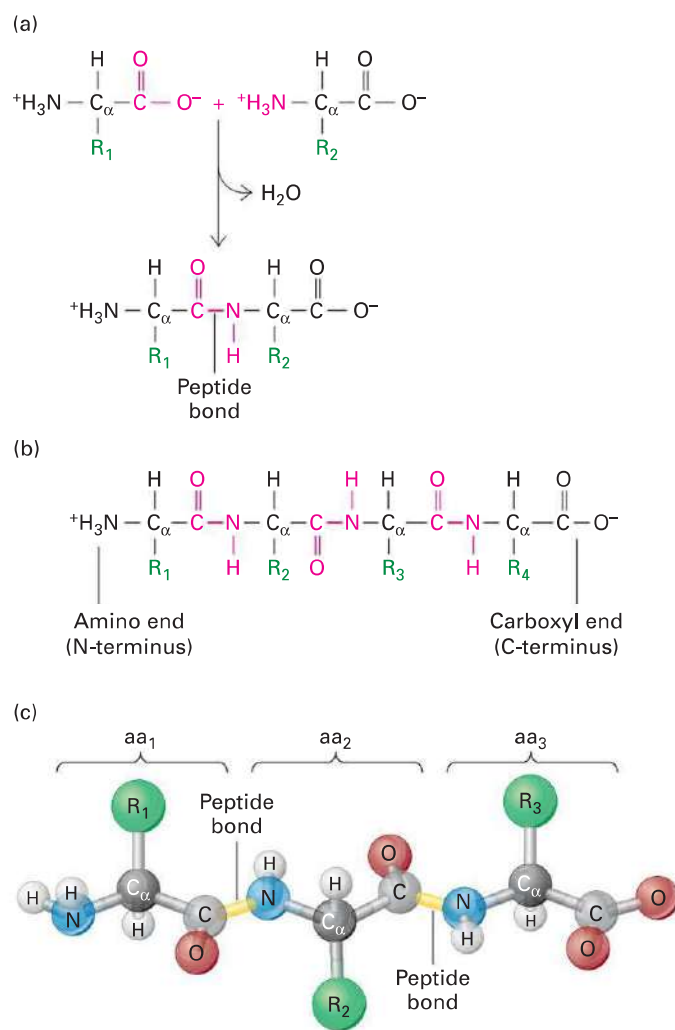
In many proteins, the polymer chain folds into a distinct three-dimensional shape that is stabilized primarily by noncovalent interactions between regions in the linear sequence of amino acids. A key concept in understanding how proteins work is that *function is often derived from three-dimensional structure, and three-dimensional structure is determined by both a protein's amino acid sequence and intramolecular noncovalent interactions*. The principles relating biological structure and function were initially formulated by the biologists Johann von Goethe (1749–1832), Ernst Haeckel (1834–1919), and D'Arcy Thompson (1860–1948), whose work has been widely influential in biology and beyond. Indeed, their ideas greatly influenced the school of “organic” architecture pioneered in the early twentieth century that is epitomized by the dicta “form follows function” (Louis Sullivan) and “form is function” (Frank Lloyd Wright). Here we consider the architecture of proteins at four levels of organization: primary, secondary, tertiary, and quaternary (Figure 3-2).



**FIGURE 3-2 Four levels of protein hierarchy.** (a) The linear sequence of amino acids linked together by peptide bonds is the primary structure. (b) Folding of the polypeptide chain into local α helices or β sheets represents secondary structure. (c) Secondary structural elements, together with various loops and turns in a single polypeptide chain, pack into a larger, independently stable tertiary structure, which may include distinct domains. (d) Some proteins consist of more than one polypeptide associated together in a quaternary structure.

### The Primary Structure of a Protein Is Its Linear Arrangement of Amino Acids

As discussed in Chapter 2, proteins are polymers constructed out of 20 different types of amino acids. Individual amino acids are linked together in linear, unbranched chains by covalent amide bonds, called **peptide bonds**. Peptide bond formation between the amino group of one amino acid and the carboxyl group of another results in the net release of a water molecule and thus is a form of dehydration reaction (Figure 3-3a). The repeated amide N, α carbon ( $C_\alpha$ ), carbonyl C, and oxygen atoms of each amino acid residue form the backbone of a protein molecule from which the various side-chain groups project (Figure 3-3b, c). As a consequence of



**FIGURE 3-3 Structure of a polypeptide.** (a) Individual amino acids are linked together by peptide bonds, which form via reactions that result in a loss of water (dehydration).  $R_1$ ,  $R_2$ , etc., represent the side chains (“R groups”) of amino acids. (b) Linear polymers of peptide-bond-linked amino acids are called *polypeptides*, which have a free amino end (N-terminus) and a free carboxyl end (C-terminus). (c) A ball-and-stick model shows peptide bonds (yellow) linking the amino nitrogen atom (blue) of one amino acid (aa) with the carbonyl carbon atom (gray) of an adjacent one in the chain. The R groups (green) extend from the α carbon atoms (black) of the amino acids. These side chains largely determine the distinct properties of individual proteins.



the peptide linkage, the backbone exhibits directionality, usually referred to as an N-to-C orientation, because all the amino groups are located on the same side of the C<sub>α</sub> atoms. Thus one end of a protein has a free (unlinked) amino group (the *N-terminus*), and the other end has a free carboxyl group (the *C-terminus*). The sequence of a protein chain is conventionally written with its N-terminal amino acid on the left and its C-terminal amino acid on the right, and the amino acids are numbered sequentially starting from the N-terminus.

The **primary structure** of a protein is simply the linear covalent arrangement, or sequence, of the amino acid residues that compose it. The first primary structure of a protein determined was that of insulin in the early 1950s. Today the number of known sequences exceeds 10 million and is growing daily. Many terms are used to denote the chains formed by the polymerization of amino acids. A short chain of amino acids linked by peptide bonds and having a defined sequence is called an **oligopeptide**, or simply a **peptide**; longer chains are referred to as **polypeptides**. Peptides generally contain fewer than 20–30 amino acid residues, whereas polypeptides are often 200–500 residues long. The longest protein described to date is the muscle protein titin, some forms of which can be more than 34,000 residues long. We generally reserve the term **protein** for a polypeptide (or complex of polypeptides) that has a well-defined three-dimensional structure.

The size of a protein or a polypeptide is expressed either as its mass in **daltons** (a dalton is 1 atomic mass unit) or as its molecular weight (MW), which is a dimensionless number equal to the mass in daltons. For example, a 10,000-MW protein has a mass of 10,000 daltons (Da), or 10 kilodaltons (kDa). Later in this chapter, we will consider different methods for measuring the sizes and other physical characteristics of proteins. The precise molecular weight of a protein that has not been covalently modified is readily determined by summing up the weights of all of its constituent amino acids as determined from its amino acid sequence. The proteins encoded by the yeast genome, for example, have an average molecular weight of 52,728 and contain, on average, 466 amino acid residues. The average molecular weight of amino acids in proteins is 113, taking into account their average relative abundances. This value can be used to estimate the number of residues in a protein of unknown sequence if you know its molecular weight or, conversely, to estimate from the number of residues in a protein its likely molecular weight. Covalent modification of one or more amino acids in a protein—for example, by phosphorylation or glycosylation (see Chapters 2 and 13)—alters the mass of those residues and thus the mass of the protein in which they reside.

How many proteins are there in a typical eukaryotic (nucleated) cell? Let's do a simple calculation for one such cell, a hepatocyte (a major type of cell in the mammalian liver). This type of cell, roughly a cube 15 μm (0.0015 cm) on a side, has a volume of  $3.4 \times 10^{-9}$  cm<sup>3</sup> (or milliliters, ml). Assuming a cell density of 1.03 g/ml, the cell would weigh  $3.5 \times 10^{-9}$  g. Since protein accounts for approximately 20 percent

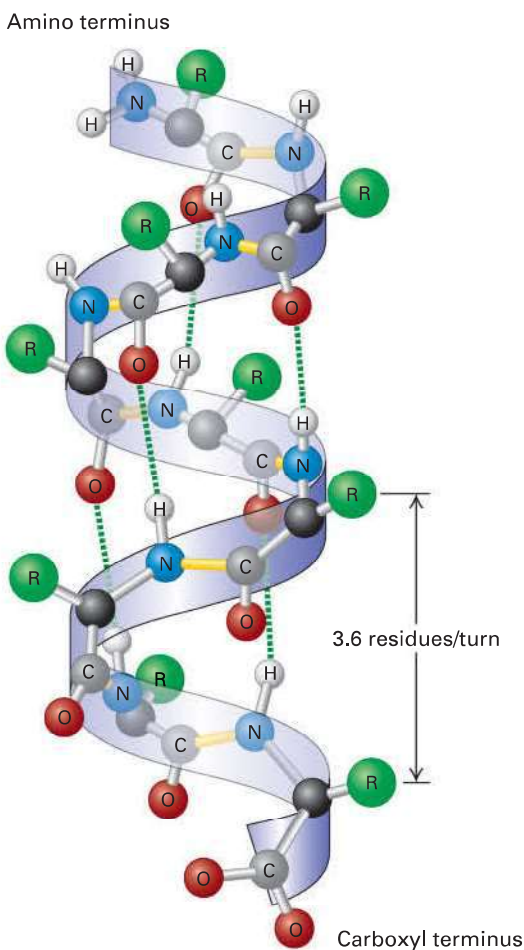
of a cell's weight, the total weight of cellular protein is  $7 \times 10^{-10}$  g. Assuming that an average protein has a molecular weight of 52,728 g/mol, we can calculate the total number of protein molecules per hepatocyte as about  $7.9 \times 10^9$  from the total protein weight and Avogadro's number, the number of molecules per mole of any chemical compound ( $6.02 \times 10^{23}$ ). To carry this calculation one step further, consider that a hepatocyte contains about 10,000 different proteins; thus each cell, on average, would contain close to a million molecules of each type of protein. In fact, the abundances of different proteins vary widely, from the quite rare insulin-binding receptor protein (20,000 molecules per cell) to the structural protein actin ( $5 \times 10^8$  molecules per cell). Every cell closely regulates the abundance of each protein such that each is present in the appropriate quantity for its cellular functions at any given time. We will learn more about the mechanisms used by cells to regulate protein levels later in this chapter and in Chapters 9 and 10.

## Secondary Structures Are the Core Elements of Protein Architecture

The second level in the hierarchy of protein structure is **secondary structure**. Secondary structures are stable spatial arrangements of segments of a polypeptide chain held together by hydrogen bonds between backbone amide and carbonyl groups and often involving repeating structural patterns. The propensity of a segment of a polypeptide chain to form any given secondary structure depends on its amino acid sequence (see Section 3.2 below). A single polypeptide may contain multiple types of secondary structure in various portions of the chain, depending on its sequence. The principal secondary structures are the **alpha (α) helix**, the **beta (β) sheet**, and the short U-shaped **beta (β) turn**. Parts of the polypeptide that don't form these structures but nevertheless have a well-defined, stable shape are said to have an *irregular* structure. The term *random coil* applies to highly flexible parts of a polypeptide chain that have no fixed three-dimensional structure. In an average protein, 60 percent of the polypeptide chain exists as α helices and β sheets; the remainder of the molecule is in irregular structures, coils, and turns. Thus α helices and β sheets are the major internal supportive elements in most proteins. Here we explore the shapes of secondary structures and the forces that favor their formation. In later sections, we examine how arrays of secondary structure fold together into larger, more complex arrangements called tertiary structure.

**The α Helix** In a polypeptide segment folded into an α helix, the backbone forms a spiral structure in which the carbonyl oxygen atom of each peptide bond is hydrogen-bonded to the amide hydrogen atom of the amino acid four residues farther along the chain in the direction of the C-terminus (Figure 3-4). Within an α helix, all the backbone amino and carboxyl groups are hydrogen-bonded to one another except at the very beginning and end of the helix. This periodic arrangement of bonds confers an amino-to-carboxy-terminal





**FIGURE 3-4 The  $\alpha$  helix, a common secondary structure in proteins.** The polypeptide backbone (seen as a ribbon) is folded into a spiral that is held in place by hydrogen bonds between backbone oxygen and hydrogen atoms. Only hydrogens involved in bonding are shown. The outer surface of the helix is covered by the side-chain R groups (green).

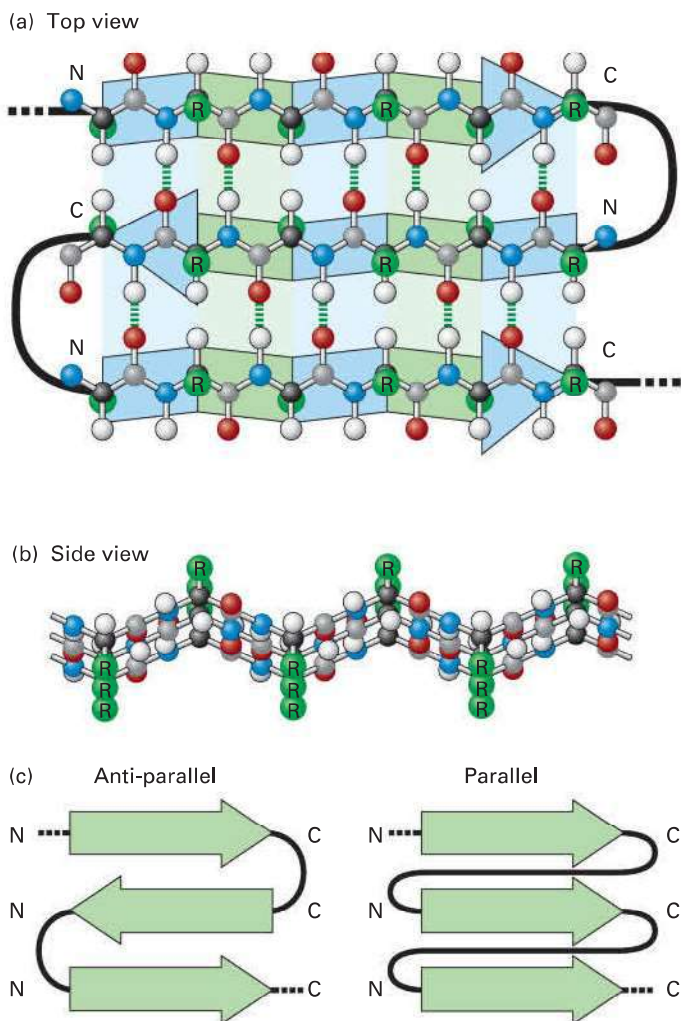
directionality on the helix because all the hydrogen bond acceptors (i.e., the carbonyl groups) have the same orientation (pointing in the downward direction in Figure 3-4), resulting in a structure in which there is a complete turn of the spiral every 3.6 residues. An  $\alpha$  helix 36 amino acids long has 10 turns of the helix and is 5.4 nm long (0.54 nm per turn).

The stable arrangement of hydrogen-bonded amino acids in the  $\alpha$  helix holds the backbone in a straight, rodlike cylinder from which the side chains point outward. The relative hydrophobic or hydrophilic quality of a particular helix within a protein is determined entirely by the characteristics of the side chains. In water-soluble proteins, hydrophilic helices with polar side chains extending outward tend to be found on the outside surfaces, where they can interact with the aqueous environment, whereas hydrophobic helices with nonpolar, hydrophobic side chains tend to be buried within the core of the folded protein. Proteins embedded in the hydrophobic core of cellular membranes (see Chapter 7)

often use one or more hydrophobic helices that are 20–25 residues long to cross the membrane. The amino acid proline is usually not found in  $\alpha$  helices because the covalent bonding of its amino group with a carbon in the side chain prevents its participation in stabilizing the backbone through normal hydrogen bonding. While the classic  $\alpha$  helix is the most intrinsically stable and most common helical form in proteins, there are variations, such as more tightly or loosely twisted helices. For example, in a specialized helix called a coiled coil (described several sections farther on), the helix is more tightly wound (3.5 residues and 0.51 nm per turn).

**The  $\beta$  Sheet** Another type of secondary structure, the  $\beta$  sheet, consists of laterally packed  $\beta$  strands. Each  $\beta$  strand is a short (5–8-residue), nearly fully extended polypeptide segment. In contrast to the  $\alpha$  helix, in which hydrogen bonds occur between the backbone amino and carboxyl groups of nearly adjacent residues, hydrogen bonds in the  $\beta$  sheet occur between backbone atoms in separate, but adjacent,  $\beta$  strands and are oriented perpendicularly to the chains of backbone atoms (Figure 3-5a). These distinct  $\beta$  strands (indicated as green and blue arrows in the figure) may be either within a single polypeptide chain, with short or long loops between the  $\beta$  strand segments, or on different polypeptide chains in a protein composed of multiple polypeptides. Figure 3-5b shows how two or more  $\beta$  strands align into adjacent rows, forming a nearly two-dimensional  $\beta$  pleated sheet (or simply *pleated sheet*), in which hydrogen bonds within the plane of the sheet hold the  $\beta$  strands together as the side chains stick out above and below the plane. Like  $\alpha$  helices,  $\beta$  strands have a directionality defined by the orientation of the peptide bonds. Therefore, in a pleated sheet, adjacent  $\beta$  strands can be oriented in alternating opposite (antiparallel) directions (see Figure 3-5a) or in the same (parallel) direction (Figure 3-5c). In some proteins,  $\beta$  sheets form part of the hydrophobic core of the protein (described below) or the side of an open space that binds other molecules; in some proteins embedded in membranes, the  $\beta$  sheets curve around and form a hydrophilic central pore through which ions and small molecules may flow (see Chapter 7).

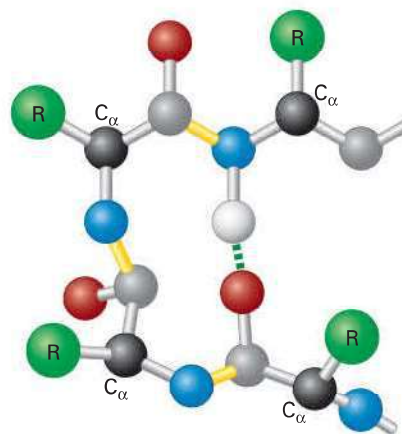
**The  $\beta$  Turn** Composed of four residues,  $\beta$  turns are located on the surface of a protein, forming sharp bends that reverse the direction of the polypeptide backbone, often toward the protein's interior. These short, U-shaped secondary structures are often stabilized by a hydrogen bond between their end residues (Figure 3-6). Glycine and proline are commonly found in  $\beta$  turns. The lack of a large side chain in glycine and the presence of a built-in bend in proline allow the polypeptide backbone to fold into a tight U shape.  $\beta$  Turns help long polypeptides fold into highly compact structures. A reversal in the direction of the polypeptide backbone may also be mediated by segments of the polypeptide that are longer than four residues and that form bends or loops. In contrast to tight  $\beta$  turns, which exhibit just a few well-defined conformations, longer loops can have many different conformations.



**FIGURE 3-5 The  $\beta$  sheet, another common secondary structure in proteins.** (a) Top view of a three-stranded  $\beta$  sheet. Each strand is highlighted by a ribbon-like arrow with alternating blue and green segments that is pointed with an N-to-C orientation, with the loops of connecting residues indicated by thick black lines. In this antiparallel  $\beta$  sheet, each strand (arrow) points in the direction opposite to that of the adjacent strand. The stabilizing hydrogen bonds between the  $\beta$  strands are indicated by green dashed lines. (b) Side view of an antiparallel  $\beta$  sheet. The projection of the R groups (green) above and below the plane of the sheet is obvious in this view. The fixed bond angles in the polypeptide backbone produce a pleated contour represented in panel (a) by the alternating colored segments. (c) Top view of two  $\beta$  sheets, whose individual strands (N-to-C orientations represented by arrows) are either antiparallel, in which the strands alternately point in opposite directions (left), or parallel, in which all strands point in the same direction (right).

## Tertiary Structure Is the Overall Folding of a Polypeptide Chain

**Tertiary structure** refers to the overall conformation of a polypeptide chain—that is, the three-dimensional arrangement of all its amino acid residues. In contrast to secondary structures, which are stabilized only by hydrogen bonds, tertiary structure is stabilized primarily by hydrophobic interactions between nonpolar side chains, together with hydrogen



**FIGURE 3-6 Structure of a  $\beta$  turn.** Composed of four residues,  $\beta$  turns reverse the direction of a polypeptide chain (resulting in a  $180^\circ$  U-turn). The  $C_\alpha$  carbons of the first and fourth residues are usually less than 0.7 nm apart, and those residues are often linked by a hydrogen bond.  $\beta$  turns facilitate the folding of long polypeptides into compact structures.

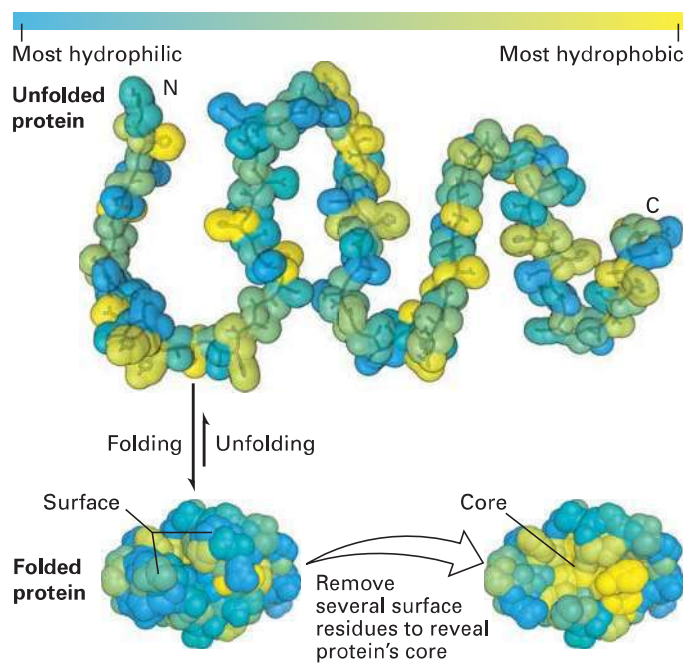
bonds involving polar side chains and backbone amino and carboxyl groups. These stabilizing forces hold together elements of secondary structure— $\alpha$  helices,  $\beta$  strands, turns, and coils. Because the stabilizing interactions are often weak, however, the tertiary structure of a protein is not rigidly fixed, but undergoes continual minute fluctuations, and some segments within the tertiary structure of a protein can be so mobile that they are considered to be disordered—that is, lacking well-defined, stable, three-dimensional structure. This variation in structure has important consequences for the function and regulation of proteins.

The chemical properties of amino acid side chains help define tertiary structure. In some proteins—for example, those that are secreted from cells or are cell-surface proteins that face the extracellular environment—*disulfide bonds* between the side chains of cysteine residues can covalently link regions of the proteins, thus restricting the proteins' flexibility and increasing the stability of their tertiary structures. Amino acids with charged hydrophilic polar side chains tend to be on the outer surfaces of proteins; by interacting with water, they help to make the proteins soluble in aqueous solutions and can form noncovalent interactions with other water-soluble molecules, including other proteins. In contrast, amino acids with hydrophobic nonpolar side chains are usually sequestered away from the water-facing surfaces of a protein, in many cases forming a water-insoluble central core. This observation led to what's known as the “oil drop model” of protein conformation because the core of a protein is relatively hydrophobic, or “oily” (Figure 3-7). Uncharged hydrophilic polar side chains are found both on the surface and in the inner core of proteins.

## There Are Four Broad Structural Categories of Proteins

Proteins usually fall into one of four broad structural categories based on their tertiary structure: *globular proteins*, *fibrous proteins*, *integral membrane proteins*, and *intrinsically disordered*





**FIGURE 3-7 The oil drop model of protein folding.** The hydrophobic and hydrophilic residues of a polypeptide chain can be distributed throughout its linear sequence as illustrated in the unfolded protein (top). The color scale denotes the most most hydrophilic residues (blue) to the most hydrophobic (yellow). When the protein folds (bottom left), hydrophilic (charged and uncharged polar) side chains will often be exposed on the protein's surface, where they can form stabilizing interactions with surrounding water and ions. In contrast, the hydrophobic residues tend to cluster together in the inner core, somewhat like drops of oil in an aqueous liquid, driven away from the aqueous surroundings by the hydrophobic effect (see Chapter 2). These core residues are more easily seen when several surface residues are removed (bottom right). [Data from M. C. Vaney et al., 1996, *Acta Crystallogr., Sect. D* **52**:505, PDB ID 193L]

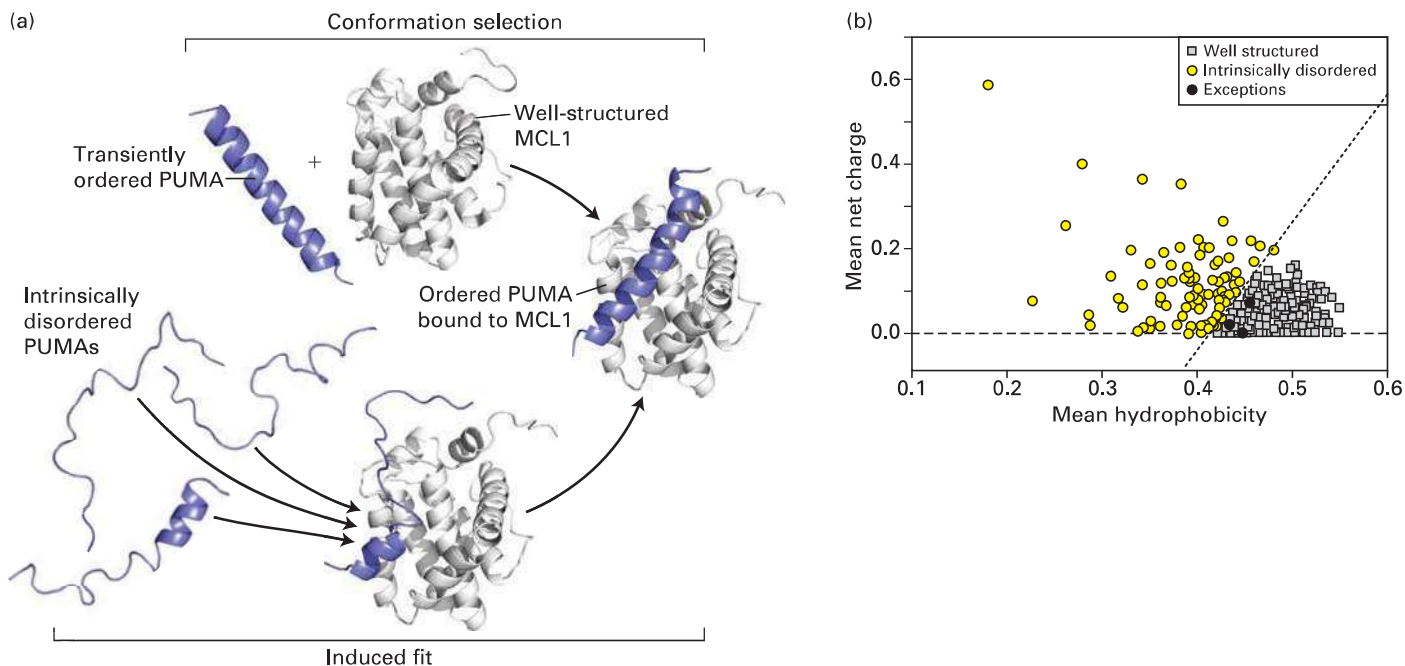
*proteins*. These four broad categories of proteins are not mutually exclusive—some proteins are made up of combinations of segments that fall into two or more of these categories. *Globular proteins* are generally water-soluble, compactly folded structures, often but not exclusively spheroidal, that comprise a mixture of secondary structures [see the structures of ras (Figure 3-9 below) and myoglobin (Figure 3-14 below)]. *Fibrous proteins* are large, elongated, often stiff molecules. Some fibrous proteins are composed of a long polypeptide chain comprising many tandem copies of a short amino acid sequence that forms a single repeating secondary structure (see the structure of collagen, the most abundant protein in mammals, in Figure 20-25). Other fibrous proteins are composed of repeating globular protein subunits, such as the helical array of G-actin protein monomers that forms F-actin microfilaments (see Chapter 17). Fibrous proteins, which often aggregate into large multiprotein fibers that do not readily dissolve in water, usually play a structural role or participate in cellular movements. *Integral membrane proteins* are embedded within the phospholipid bilayer of the membranes that enclose cells and organelles and are discussed in detail in Chapter 7.

*Intrinsically disordered proteins* are fundamentally distinct from the well-ordered proteins in the other three categories. Many proteins we consider in this book adopt only one or a few very closely related conformations when they are in their normal functional state, called the *native state*. Intrinsically disordered proteins, however, do not have well-ordered structures in their native, functional states; instead, their polypeptide chains are very flexible—indeed, disordered—with no fixed conformation. Sometimes only a segment of a polypeptide chain, rather than the entire chain, will be intrinsically disordered. The exceptional conformational flexibilities of intrinsically disordered proteins or protein segments appear to be key to their functional activities, such as the ability to interact with multiple partner proteins or to fold into a well-defined conformation only after binding to such partners (Figure 3-8a).

Intrinsically disordered proteins typically, but not exclusively, serve as signaling molecules, regulators of the activities of other molecules, or as scaffolds for multiple proteins, small molecules, and ions (e.g., binding ions via multiple charged residues). Regions of intrinsic disorder can provide flexible links, or tethers, between well-ordered regions of a protein; serve as sites of some types of post-translational protein modification [e.g., covalent addition of phosphate groups (phosphorylation) or sugars (glycosylation)]; serve as targets of protease digestion that regulates protein activity; inhibit the activity of the protein in which they are embedded (autoinhibition sites); or serve as signals for intracellular sorting of proteins (see Chapter 13). The activities of many proteins containing intrinsically disordered segments are described in subsequent chapters. For example, phosphorylation of the disordered C-terminal domain (CTD) of RNA polymerase II (see Figure 8-12), which is composed of multiple repeats of a seven-amino-acid sequence containing proline, threonine, and serine, regulates key steps in the synthesis of mRNA (see Chapters 9 and 10). The N-termini of histone proteins that control DNA organization in chromatin (see Chapter 8) are sites of important post-translational modifications, and the disordered, proline-rich FH1 region in the protein formin controls the assembly of actin filaments (see Chapter 17).

Intrinsically disordered proteins can be identified experimentally using various biochemical techniques, such as tests of sensitivity to protease digestion (disordered regions usually exhibit greater protease sensitivity), and a wide variety of biophysical techniques, including spectroscopy. The intrinsic disorder of these proteins apparently arises as a consequence of their having a sequence that, relative to well-ordered proteins, is richer in polar amino acids, proline, and net charge, and poorer in hydrophobic residues (Figure 3-8b). Algorithms primarily based on calculations of amino acid composition—particularly net charge and hydrophobicity—are used to predict which proteins or segments of proteins are intrinsically disordered. By some estimates, about 30 percent or more of eukaryotic proteins are predicted to have at least one segment of 50 or more consecutive residues that is disordered.





**EXPERIMENTAL FIGURE 3-8 Intrinsically disordered proteins: mechanisms of binding to well-ordered proteins and identification based on hydrophobicity and net charge.** (a) The binding of an intrinsically disordered protein (PUMA, blue) to a well-ordered protein (MCL1, gray) results in the formation of a well-defined structure in the previously disordered protein. PUMA and MCL1 are intracellular proteins that can influence the regulated process of cell death called apoptosis (see Chapter 21). Two mechanisms have been proposed for generating a bound complex in which both proteins are structured: conformational selection (top pathway) and induced fit (bottom pathway). In conformational selection, the disordered protein (PUMA) occasionally and transiently adopts in solution the structure it would have in the bound state. The well-ordered binding partner (MCL1) can then bind to (select) PUMA in that transient, ordered conformation, forming a relatively stable bound complex. In induced fit, the disordered protein begins to bind to the well-ordered partner while still disordered and then, while bound, is induced to form the ordered conformation present in the relatively stable, heterodimeric complex. Recent experiments suggest

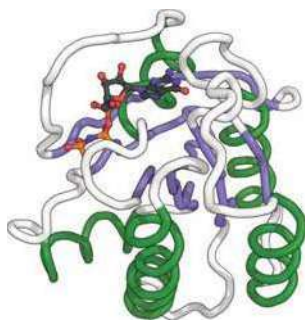
that the induced fit mechanism best describes the binding of PUMA and MCL1. (b) The sequences of 275 well-ordered, monomeric globular proteins (gray squares) and 91 intrinsically disordered proteins (black and yellow circles) were used to calculate the mean hydrophobicity per residue in each protein using a scale of 0 (least hydrophobic) to 1 (most hydrophobic, x axis), and the mean net charge per residue at pH 7.0 (y axis). With only three exceptions (black circles), the proteins define two distinct distributions: low hydrophobicity, high net charge (intrinsically disordered, yellow circles) and high hydrophobicity, low net charge (well-ordered, gray squares). The three disordered proteins (black circles) that overlap with the well-ordered population each contain substantial segments predicted to be disordered (low hydrophobicity, high net charge) that apparently overwhelm the rest of the proteins' sequences that might otherwise result in a well-ordered conformation. [Part (a) from Rogers, J. et al., "Folding and Binding of an Intrinsically Disordered Protein: Fast, but Not 'Diffusion-Limited,'" *J. Am. Chem. Soc.*, 2013, 135 (4), pp1415-1422. <http://pubs.acs.org/doi/pdf/10.1021/ja309527h>. Part (b) data from V. N. Uversky, J. R. Gillespie, and A. L. Fink, 2000, *Proteins* **41**:415-427.]

## Different Ways of Depicting the Conformation of Proteins Convey Different Types of Information

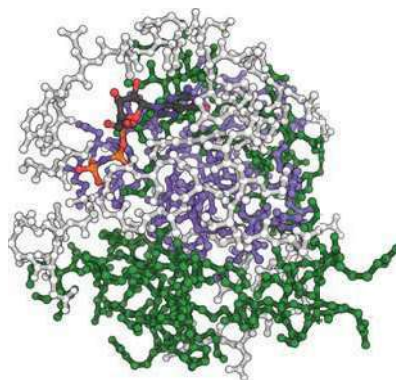
The simplest way to represent three-dimensional protein structure is to trace the course of the backbone atoms, sometimes only the  $C_{\alpha}$  atoms, with a solid line (called a  $C_{\alpha}$  backbone trace, Figure 3-9a); the most complex representation, called a ball-and-stick model, shows every atom (Figure 3-9b). The  $C_{\alpha}$  backbone trace shows the overall folding of the polypeptide chain without consideration of the amino acid side chains; the ball-and-stick model (with balls representing atoms and sticks representing bonds) details the interactions between side-chain atoms, including those that stabilize the protein's conformation and interact with other molecules, as well as the atoms of the backbone. Even though both views are useful, the elements of secondary structure are not always easily discerned in them. Another type of representation, called a ribbon diagram, uses common shorthand symbols for depicting secondary structure—for example,

coiled ribbons or solid cylinders for  $\alpha$  helices, flat ribbons or arrows for  $\beta$  strands, and flexible thin strands for  $\beta$  turns, coils, and loops (Figure 3-9c). In a variation of the basic ribbon diagram, ball-and-stick or space-filling models of all or only a subset of side chains can be attached to the backbone ribbon. In this way, side chains that are of interest can be visualized in the context of the secondary structure that is especially clearly represented by the ribbons.

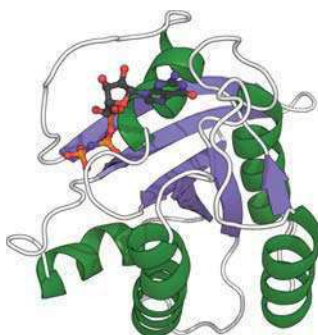
However, none of these three ways of representing protein structure conveys much information about the atoms that are on the protein's surface and in contact with the watery environment. The surface is of interest because it is where other molecules usually bind to a protein. Thus a useful alternative way to represent proteins is to show only the water-accessible surface and use colors to highlight regions having a common chemical character, such as hydrophobicity or hydrophilicity, and charge characteristics, such as positive (basic) or negative (acidic) side chains (Figure 3-9d). Such models reveal the topography of the protein surface

(a) C<sub>α</sub> backbone trace

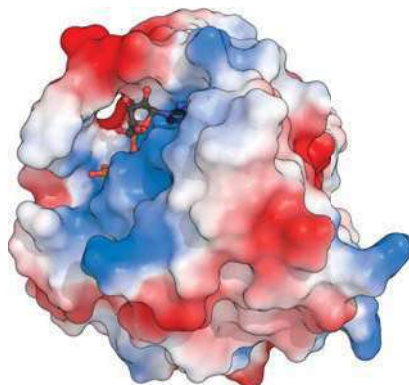
(b) Ball-and-stick model



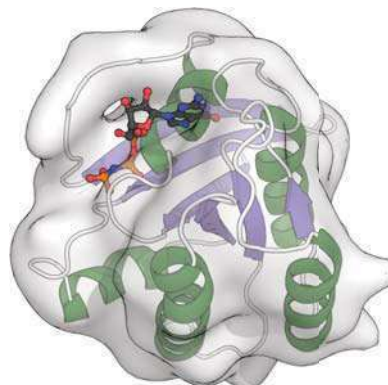
(c) Ribbon diagram



(d) Water-accessible surface



(e) Hybrid model



**FIGURE 3-9 Five ways to visualize the protein Ras with its bound GDP.** (a) The C<sub>α</sub> backbone trace demonstrates how the polypeptide is tightly packed into a small volume. (b) A ball-and-stick representation reveals the locations of all atoms. (c) Turns and loops connect pairs of helices and strands. (d) A water-accessible surface reveals the numerous lumps, bumps, and crevices on the protein surface. Regions of positive charge are shaded purple; regions of negative charge are shaded red. (e) Hybrid model in which ribbon and transparent surface models are combined. [Data from E. F. Pai et al., 1990, *EMBO J.* **9**:2351–2359, PDB ID 5p21.]

and the distribution of charge, both important features of binding sites, as well as clefts in the surface where other molecules may bind. This view represents a protein as it is “seen” by another molecule.

### Structural Motifs Are Regular Combinations of Secondary Structures

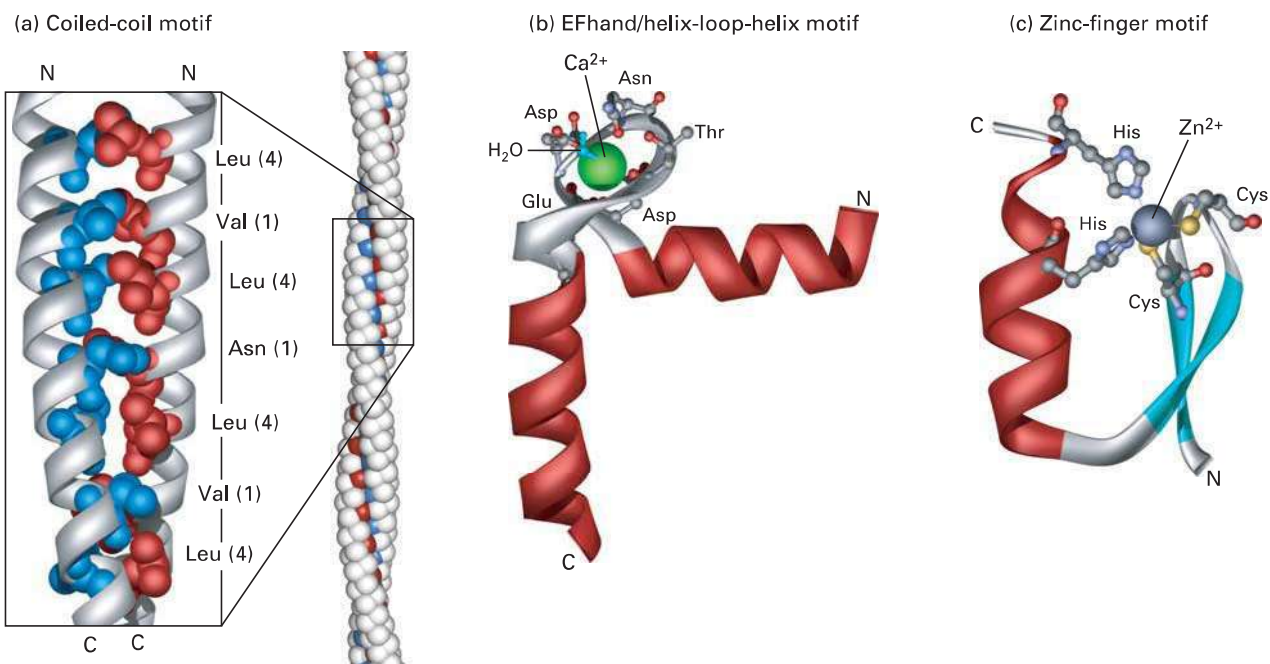
A particular combination of two or more secondary structures that form a distinct three-dimensional structure is called a **structural motif** when it appears in multiple proteins. A structural motif is often, but not always, associated with a specific function. Any particular structural motif will frequently perform a common function in different proteins, such as binding to a particular ion or small molecule—for example, calcium or ATP. Some structural motifs, when isolated from the rest of a protein, are stable, and are thus called *structural domains*, as we shall see shortly. However other structural motifs do not form thermodynamically stable structures in the absence of other portions of the protein and are thus not considered to be independent structural domains.

One common structural motif is the  $\alpha$  helix–based **coiled coil**, or heptad repeat. Many proteins, including fibrous proteins and DNA-regulating proteins called transcription factors (see Chapter 9), assemble into dimers or trimers by using a coiled-coil motif, in which  $\alpha$  helices from two, three, or even four separate polypeptide chains coil about one another—resulting in a coil of coils; hence the name (Figure 3-10a). The individual helices bind tightly to one

another because each helix has a strip of aliphatic (hydrophobic, but not aromatic) side chains (leucine, valine, etc.) running along one side of the helix that interacts with a similar strip in the adjacent helix, thus sequestering the hydrophobic groups away from water and stabilizing the assembly of multiple independent helices. These hydrophobic strips are generated along only one side of the helix because the primary structure of each helix is composed of repeating seven-amino-acid units, called **heptads**, in which the side chains of the first and fourth residues are aliphatic and the other side chains are often hydrophilic (see Figure 3-10a). Because hydrophilic side chains extend from one side of the helix and hydrophobic side chains extend from the opposite side, the overall helical structure is **amphipathic**. Because leucine frequently appears in the fourth positions and the hydrophobic side chains merge together like the teeth of a zipper, these structural motifs are also called **leucine zippers**.

Many other structural motifs contain  $\alpha$  helices. A common calcium-binding motif called the **EF hand** contains two short helices connected by a loop (Figure 3-10b). This structural motif, one of several **helix-turn-helix** and **helix-loop-helix** structural motifs, is found in more than a hundred proteins and is used for sensing calcium levels. The binding of a Ca<sup>2+</sup> ion to oxygen atoms in conserved residues in the loop depends on the concentration of Ca<sup>2+</sup> in the cell and sometimes induces a conformational change in the protein, altering its activity. Thus calcium concentrations can directly control proteins’ structures and functions. Somewhat different helix-turn-helix and **basic helix-loop-helix**





**FIGURE 3-10 Motifs of protein secondary structure.** (a) This parallel two-stranded coiled-coil motif (*left*) is characterized by two  $\alpha$  helices wound around each other. Helix packing is stabilized by interactions between hydrophobic side chains (red and blue) present at regular intervals along each strand and found along the seam of the intertwined helices. Each  $\alpha$  helix exhibits a characteristic heptad repeat sequence with a hydrophobic residue often, but not always, at positions 1 and 4, as indicated. The coiled-coil nature of this structural motif is more apparent in long coiled coils containing many such motifs (*right*). (b) An EF hand, a type of helix-loop-helix motif, consists of two helices connected by a short loop in a specific conformation. This structural motif is common to many proteins, including many calcium-binding and DNA-binding regulatory proteins.

In calcium-binding proteins such as calmodulin, oxygen atoms from five residues in the acidic glutamate- and aspartate-rich loop and one water molecule form ionic bonds with a  $\text{Ca}^{2+}$  ion. (c) The zinc-finger motif is present in many DNA-binding proteins that help regulate transcription. A  $\text{Zn}^{2+}$  ion is held between a pair of  $\beta$  strands (blue) and a single  $\alpha$  helix (red) by a pair of cysteine residues and a pair of histidine residues. The two invariant cysteine residues are usually at positions 3 and 6, and the two invariant histidine residues are at positions 20 and 24 in this 25-residue motif. [Part (a) data from L. Gonzalez, Jr., D. N. Woolfson, and T. Alber, 1996, *Nat. Struct. Biol.* **3**:1011–1018, PDB IDs 1zik and 2tma. Part (b) data from R. Chattopadhyaya et al., 1992, *J. Mol. Biol.* **228**:1177–1192, PDB ID 1cll. Part (c) data from S. A. Wolfe, R. A. Grant, and C. O. Pabo, 2003, *Biochemistry* **42**:13401–13409, PDB ID 1llm.]

(bHLH) structural motifs are used for protein binding to DNA and, consequently, for the regulation of gene activity (see Chapter 9). Yet another structural motif commonly found in proteins that bind RNA or DNA is the **zinc finger**, which contains three secondary structures—an  $\alpha$  helix and two  $\beta$  strands with an antiparallel orientation—that form a fingerlike bundle held together by a zinc ion (Figure 3-10c).

The relationship between the primary structure of a polypeptide chain and the structural motifs into which it folds is not always straightforward. The amino acid sequences responsible for any given structural motif in different proteins may be very similar to one another. In other words, a common *sequence motif* can result in a common structural motif. This is the case for the heptad repeats that form coiled coils. However, it is also possible for seemingly unrelated amino acid sequences to fold into a common structural motif, so it is not always possible to predict which amino acid sequences will fold into a given structural motif. Conversely, it is possible that a commonly occurring sequence motif will not fold into a well-defined structural motif. Sometimes short sequence motifs that have an unusual abundance of a particular amino acid, such as proline or aspartate or glutamate, are called “domains”; however, these

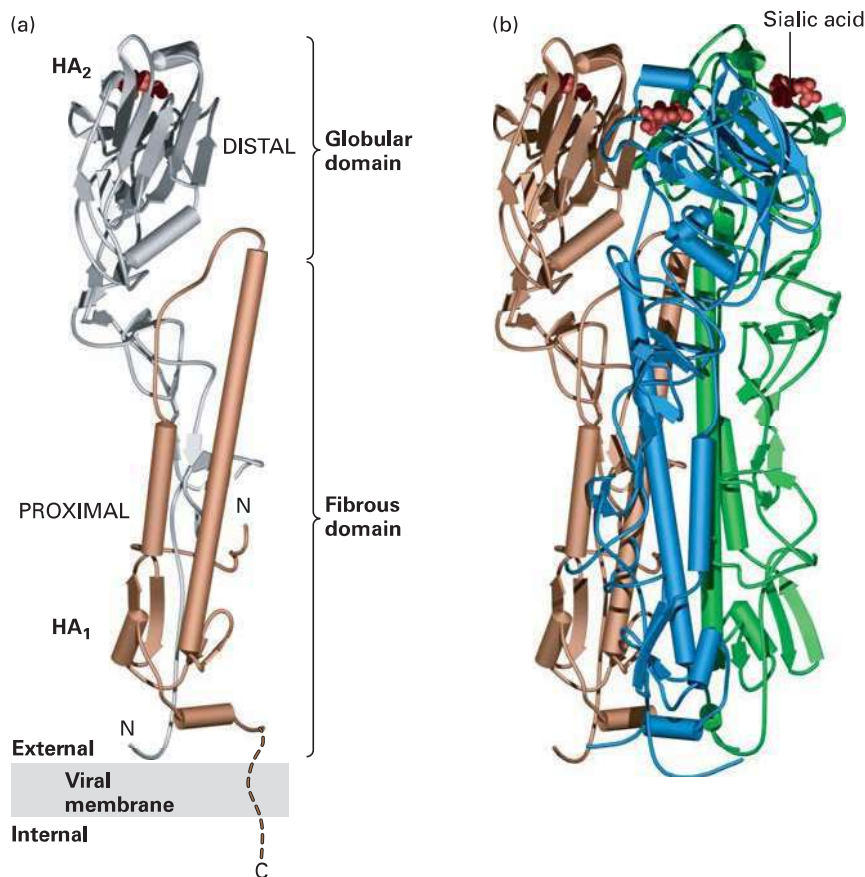
and other short contiguous segments are more appropriately called “sequence motifs” than “domains,” as the latter term has a distinct meaning that we will define shortly.

We will encounter numerous additional motifs in our discussions of proteins in this and other chapters. The presence of the same structural motif in different proteins with similar functions clearly indicates that these useful combinations of secondary structures have been conserved in evolution.

## Domains Are Modules of Tertiary Structure

Distinct regions of protein structure are often referred to as **domains**. There are three main classes of protein domains: functional, structural, and topological. A *functional domain* is a region of a protein that exhibits a particular activity characteristic of that protein, usually even when isolated from the rest of the protein. For instance, a particular region of a protein may be responsible for its catalytic activity (e.g., a kinase domain that covalently adds a phosphate group to another molecule) or its binding ability (e.g., a DNA-binding domain or a membrane-binding domain). Functional domains are often identified experimentally by whittling down a protein to its smallest active fragment with the aid of **proteases**,





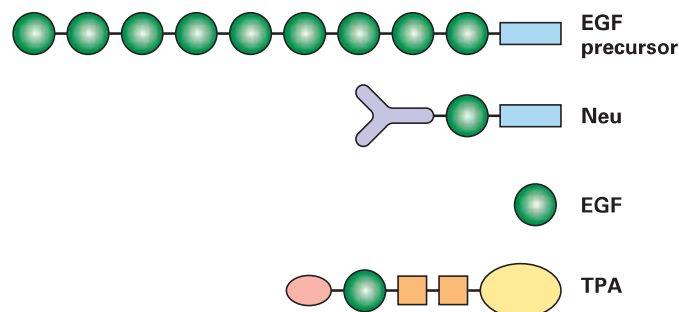
**FIGURE 3-11 Tertiary and quaternary levels of structure.** The protein pictured here, hemagglutinin (HA), is found on the surface of the influenza virus. This long multimeric molecule has three identical subunits, each composed of two polypeptide chains, HA<sub>1</sub> and HA<sub>2</sub>. (a) The tertiary structure of each HA subunit comprises the folding of its helices and strands into a compact structure that is 13.5 nm long and divided into two domains. The membrane-distal domain (silver) is folded into a globular conformation. The membrane-proximal domain (gold) has a fibrous, stemlike conformation owing to the alignment of two long  $\alpha$  helices (cylinders) of HA<sub>2</sub> with  $\beta$  strands in HA<sub>1</sub>. Short turns and longer loops, many of them at the surface of the molecule, connect the helices and strands in each chain. (b) The quaternary structure of HA is stabilized by lateral interactions between the long helices (cylinders) in the fibrous domains of the three subunits (gold, blue, and green), forming a triple-stranded coiled-coil stalk. Each of the distal globular domains in HA binds sialic acid (red) on the surface of target cells. Like many membrane proteins, HA contains several covalently linked carbohydrate chains (not shown). [Data from S. J. Gamblin et al., 2004, *Science* **303**:1838–1842, PDB ID 1ruz.]

enzymes that cleave one or more peptide bonds in a target polypeptide. Alternatively, the DNA encoding a protein can be modified so that when the modified DNA is used to generate a protein, only a particular region, or domain, of the full-length protein is made. Thus it is possible to determine if specific parts of a protein are responsible for particular activities exhibited by the protein. Indeed, functional domains are often also associated with corresponding structural domains.

A *structural domain* is a region about 40 or more amino acids in length, arranged in a single, stable, and distinct structure often comprising one or more secondary structures. Many structural domains can fold into their characteristic structures independently of the rest of the protein in which they are embedded. As a consequence, distinct structural domains can be linked together—sometimes by short or long spacers—to form a large multidomain protein. Each of the polypeptide chains in the trimeric flu virus hemagglutinin, for example, contains a globular domain and a fibrous domain (Figure 3-11a). Structural domains can be incorporated as modules into different proteins. The modular approach to protein architecture is particularly easy to recognize in large proteins, which tend to be mosaics of different domains that confer distinct activities and thus can perform different functions simultaneously. As many as 75 percent of the proteins in eukaryotes have multiple structural domains. Structural domains frequently are also functional domains in that they can have an activity independent of the rest of the protein.

The epidermal growth factor (EGF) domain is a structural domain that is present in several proteins (Figure 3-12). EGF

is a small, soluble peptide hormone that binds to cells in the embryo and in skin and connective tissue in adults, causing them to divide. It is generated by proteolytic cleavage (breaking of a peptide bond) between repeated EGF domains in the EGF precursor protein, which is anchored in the plasma membrane by a membrane-spanning domain. EGF domains with sequences similar to, but not identical to, that of the EGF peptide hormone are present in other proteins and can be liberated by proteolysis. These proteins include tissue plasminogen activator (TPA), a protease that is used to dissolve blood



**FIGURE 3-12 Modular nature of protein domains.** Epidermal growth factor (EGF) is generated by proteolytic cleavage of a precursor protein containing multiple EGF domains (green) and a membrane-spanning domain (blue). An EGF domain is also present in the Neu protein and in tissue plasminogen activator (TPA). These proteins also contain other widely distributed domains, indicated by shape and color. See I. D. Campbell and P. Bork, 1993, *Curr. Opin. Struc. Biol.* **3**:385.

clots in heart attack victims; Neu protein, which takes part in embryonic differentiation; and Notch protein, a receptor protein in the plasma membrane that functions in developmentally important signaling (see Chapter 16). Besides the EGF domain, these proteins have other domains in common with other proteins. For example, TPA possesses a trypsin domain, a functional domain found in some proteases. It is estimated that there are about a thousand different types of structural domains in all proteins. Some of these are not very common, whereas others are found in many different proteins. Indeed, by some estimates, only nine major types of structural domains account for as much as a third of all the structural domains in all proteins. Structural domains can be recognized in proteins whose structures have been determined by x-ray crystallography or nuclear magnetic resonance (NMR) analysis or in images captured by electron microscopy.

Regions of proteins that are defined by their distinctive spatial relationships to the rest of the protein are *topological domains*. For example, some proteins associated with cell-surface membranes have a part extending inward into the cytoplasm (cytoplasmic domain), a part embedded within the phospholipid bilayer (membrane-spanning domain), and a part extending outward into the extracellular space (extracellular domain). Each of these parts can comprise one or more structural and functional domains.

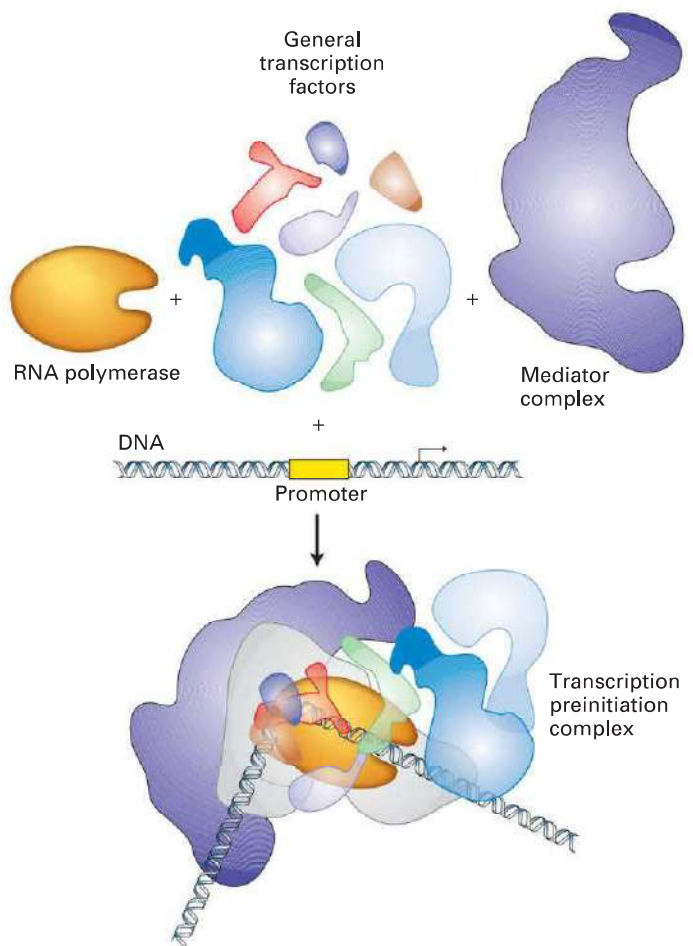
In Chapter 8, we will consider the mechanism by which the gene segments that correspond to domains became shuffled in the course of evolution, resulting in their appearance in many proteins. Once a functional, structural, or topological domain has been identified and characterized in one protein, it is possible to use that information to search for similar domains in other proteins and to suggest potentially similar functions for those domains in those proteins.

## Multiple Polypeptides Assemble into Quaternary Structures and Supramolecular Complexes

Multimeric proteins consist of two or more polypeptide chains, which in this context are referred to as *subunits*. A fourth level of structural organization, **quaternary structure**, describes the number (stoichiometry) and relative positions of the subunits in multimeric proteins (Figure 3-2). Flu virus hemagglutinin, for example, is a trimer of three identical subunits (a homotrimer) held together by noncovalent bonds (Figure 3-11b). Other multimeric proteins are composed of various numbers of identical (homomeric) or different (heteromeric) subunits. Hemoglobin, the oxygen-carrying molecule in blood, is an example of a heteromeric multimeric protein, as it has two copies each of two different polypeptide chains (as discussed below). In many cases, the individual monomer subunits of a multimeric protein cannot function normally unless they are assembled into the multimeric protein. In other cases, assembly into a multimeric protein permits proteins that act sequentially in a pathway to increase their efficiency of operation owing to their juxtaposition in space, a phenomenon referred to as *metabolic coupling*. Classic examples of metabolic coupling are the fatty acid synthases, the enzymes in fungi that synthesize fatty acids, and the polyketide synthases,

the large multiprotein complexes in bacteria that synthesize a diverse set of pharmacologically relevant molecules called polyketides, including the antibiotic erythromycin.

The highest level in the hierarchy of protein structure is the association of proteins into supramolecular complexes. Typically, such structures are very large, in some cases exceeding 1 megadalton (MDa) in mass, approaching 30–300 nm in size, and containing tens to hundreds of polypeptide chains and sometimes other biopolymers such as nucleic acids. The capsid that encases the nucleic acids of the viral genome is an example of a supramolecular complex with a structural function. The bundles of cytoskeletal filaments that support and give shape to the plasma membrane are another example. Other supramolecular complexes act as molecular machines, carrying out the most complex cellular processes by integrating multiple proteins, each with distinct functions, into one large assembly. For example, a transcriptional machine is responsible for synthesizing messenger RNA (mRNA) using a DNA template. This transcriptional



**FIGURE 3-13 A molecular machine: the transcription initiation complex.** The core RNA polymerase, general transcription factors, a mediator complex containing about 20 subunits, and other protein complexes not depicted here assemble at a promoter in DNA. The polymerase carries out transcription of DNA; the associated proteins are required for initial binding of the polymerase to a specific promoter. The multiple components function together as a molecular machine.

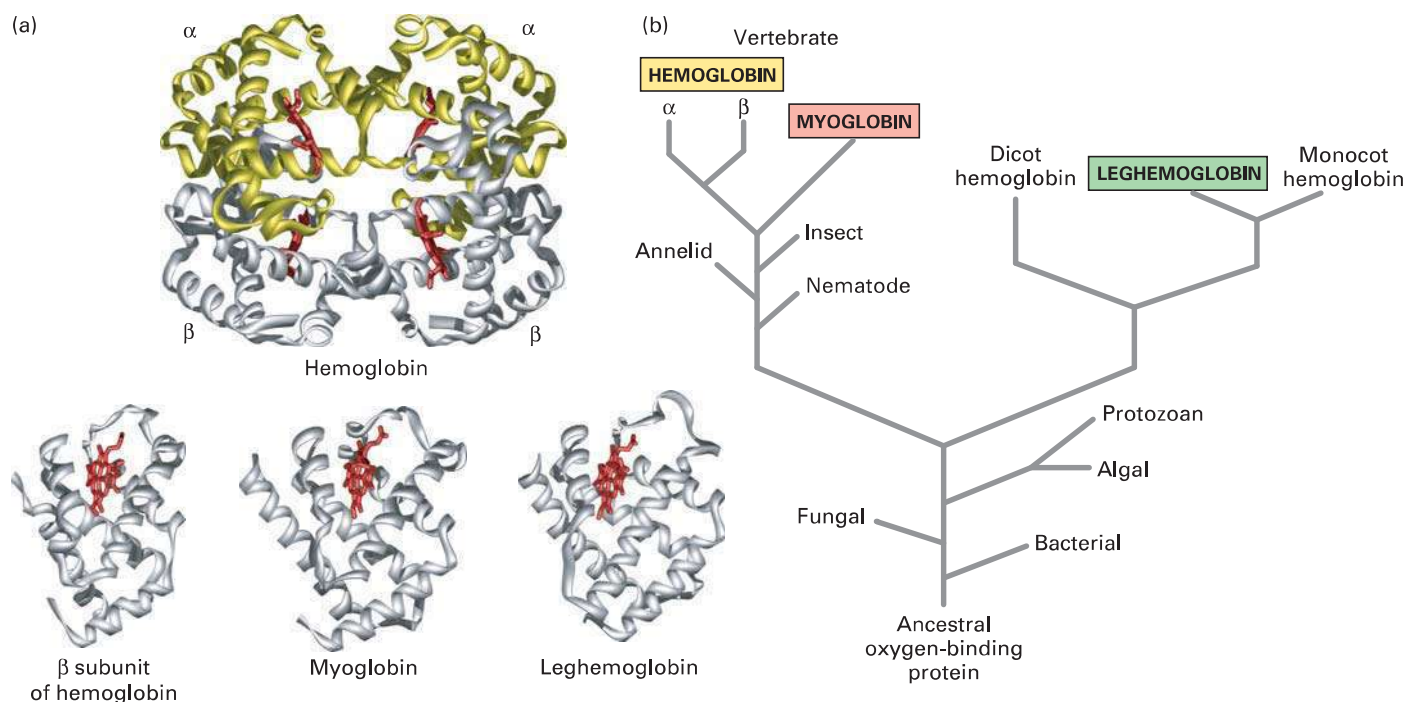
machine, the operational details of which are discussed in Chapters 5 and 9, consists of RNA polymerase, itself a multimeric protein, and at least 50 additional components, including general transcription factors, promoter-binding proteins, helicase, and other protein complexes (Figure 3-13). Ribosomes, also discussed in Chapter 5, are complex multiprotein and multi-nucleic acid machines that synthesize proteins. One of the most complex multiprotein assemblies is the nuclear pore, a structure that allows communication and passage of macromolecules between the nucleus and the cytoplasm (see Chapter 14). It is composed of multiple copies of about 30 distinct proteins and forms an assembly with an estimated mass of 50 MDa. The fatty acid synthases and polyketide synthases referred to above are also molecular machines.

### Comparing Protein Sequences and Structures Provides Insight into Protein Function and Evolution

Analyses of many diverse proteins have conclusively established a relation between the amino acid sequence, three-dimensional structure, and function of proteins. One of the earliest examples involved a comparison of two oxygen-carrying proteins: myoglobin in muscle and hemoglobin in red blood cells. Myoglobin—a monomer (consisting of one polypeptide chain/protein molecule)—and hemoglobin—a

tetramer (consisting of two  $\alpha$  and two  $\beta$  polypeptides, or subunits, per protein)—both contain a heme group noncovalently attached to each polypeptide chain (Figure 3-14a). The heme group binds oxygen. A mutation in the gene encoding the  $\beta$  chain of hemoglobin that results in the substitution of a valine for a glutamic acid disturbs this protein's folding and function and causes sickle-cell disease (also called sickle-cell anemia). The properly aligned sequences of the 141-residue myoglobin and the 153-residue  $\beta$  subunit of hemoglobin have 40 residues in equivalent positions in the sequences that are identical and another 21 that have side chains that are chemically very similar. This high degree of identity and similarity (43 percent of the myoglobin residues) is consistent with their similar oxygen-binding functions. X-ray crystallographic analysis showed that the three-dimensional structures of myoglobin and of the  $\alpha$  and  $\beta$  subunits of hemoglobin, as well as that of the evolutionarily distant oxygen-carrying leghemoglobin from plants, are remarkably similar (see Figure 3-14a).

A good rule of thumb is that the greater the similarity of the sequences of two polypeptide chains, the more likely they are to have similar three-dimensional structures and similar functions. While this comparative approach is very powerful, caution must always be exercised when attributing to one protein, or a part of a protein, a function or structure similar to that of another protein based only on amino acid sequence



**FIGURE 3-14 Evolution of the globin protein family.**

(a) Hemoglobin is a tetramer of two  $\alpha$  and two  $\beta$  subunits. The structural similarity of these subunits to leghemoglobin and myoglobin, both of which are monomers, is evident. A heme molecule (red) noncovalently associated with each globin polypeptide is directly responsible for oxygen binding in these proteins. (b) A primitive monomeric oxygen-binding globin is thought to be the ancestor of modern-day blood hemoglobins, muscle myoglobins, and plant leghemoglobins. Sequence comparisons have revealed that the evolution of the globin

proteins parallels the evolution of animals and plants. Major changes occurred with the divergence of plant globins from animal globins and of myoglobin from hemoglobin. Later, gene duplication gave rise to the  $\alpha$  and  $\beta$  subunits of hemoglobin. See R. C. Hardison, 1996, *P. Natl. Acad. Sci. USA* **93**:5675. [Part (a) data from G. Fermi et al., 1984, *J. Mol. Biol.* **175**:159–174, PDB ID 2hbb (hemoglobin), H. C. Watson, 1969, *Prog. Stereochem.* **4**:299, PDB ID 1mbn (myoglobin), and M. S. Hargrove et al., 1997, *J. Mol. Biol.* **266**:1032–1042, PDB ID 1bin (leghemoglobin).]



similarities. There are examples in which proteins with similar overall structures display different functions, as well as cases in which functionally unrelated proteins with dissimilar amino acid sequences nevertheless have very similar folded tertiary structures, as will be explained below. Nevertheless, in many cases, such comparisons of sequences provide important insights into protein structure and function.

Use of sequence comparisons to deduce protein structure and function has expanded substantially in recent years as the genomes and messenger RNAs of more and more organisms have been sequenced, permitting a vast array of protein sequences to be deduced. Indeed, the molecular revolution in biology during the last decades of the twentieth century created a new scheme of biological classification based on similarities and differences in the amino acid sequences of proteins. Proteins that have a common ancestor are referred to as **homologs**. The main evidence for **homology** among proteins, and hence for their common ancestry, is similarity in their sequences, which is often reflected in similar structures. We can describe homologous proteins as belonging to a “family” and can trace their lineage—how closely or distantly they are related to one another in an evolutionary sense—from comparisons of their sequences. Generally, more closely related proteins exhibit greater sequence similarity than more distantly related proteins because, over evolutionary time, mutations accumulate in the genes encoding these proteins. The folded three-dimensional structures of homologous proteins may be similar even if some parts of their primary structure show little evidence of sequence homology. Initially, proteins with relatively high sequence similarities (>50 percent exact amino acid matches, or “identities”) and related functions or structures were defined as an evolutionarily related *family*, while a *superfamily* encompassed two or more families in which the interfamily sequences matched less well (~30–40 percent identities) than within one family. It is generally thought that proteins with about 30 percent sequence identity are likely to have similar three-dimensional structures; however, such high sequence identity is not required for proteins to share similar structures. Revised definitions of *family* and *superfamily* have been proposed, in which a family comprises proteins with a clear evolutionary relationship (>30 percent identity or additional structural and functional information showing common descent but <30 percent identity), while a superfamily comprises proteins with only a probable common evolutionary origin—for example, lower sequence identities but one or more common motifs or domains.

The kinship among homologous proteins is most easily visualized by a tree diagram based on sequence analyses. For example, the amino acid sequences of globins—the proteins hemoglobin and myoglobin and their relatives from bacteria, plants, and animals—suggest that they evolved from an ancestral monomeric oxygen-binding protein (Figure 3-14b). With the passage of time, the gene for this ancestral protein slowly changed, initially diverging into lineages leading to animal and plant globins. Subsequent changes gave rise to myoglobin and to the  $\alpha$  and  $\beta$  subunits of the tetrameric hemoglobin molecule ( $\alpha_2\beta_2$ ) of the vertebrate circulatory system.

## KEY CONCEPTS OF SECTION 3.1

### Hierarchical Structure of Proteins

- Proteins are linear polymers of amino acids linked together by peptide bonds. A protein can have a single polypeptide chain or multiple polypeptide chains. The primary structure of a polypeptide chain is the sequence of covalently linked amino acids that compose the chain. Various, mostly noncovalent interactions between amino acids in the linear sequence stabilize a protein's specific folded three-dimensional structure, or conformation.
- The  $\alpha$  helix,  $\beta$  strand and sheet, and  $\beta$  turn are the most prevalent elements of protein secondary structure. Secondary structures are stabilized by hydrogen bonds between atoms of the peptide backbone (see Figures 3-4–3-6).
- Protein tertiary structure results from hydrophobic interactions between nonpolar side groups and from hydrogen bonds and ionic interactions involving polar side groups and the polypeptide backbone. These interactions stabilize the folding of the protein, including its secondary structural elements, into an overall three-dimensional arrangement.
- Entire proteins or segments of proteins usually fall into one of four broad structural categories: globular proteins, fibrous proteins, integral membrane proteins, and intrinsically disordered proteins.
- The exceptional conformational flexibilities of intrinsically disordered proteins contribute to their functions as binding partners, signaling molecules, regulators of other molecules, scaffolds, flexible links between well-ordered regions of a protein, sites of post-translational protein modification, autoinhibitors, and signals for intracellular protein sorting.
- Certain combinations of secondary structures give rise to structural motifs, which are found in a variety of proteins and are often associated with specific functions (see Figure 3-10).
- Proteins often contain distinct domains, independently folded regions with characteristic structural, functional, and/or topological properties.
- The incorporation of domains as modules in different proteins in the course of evolution has generated diversity in protein structure and function.
- The number and organization of individual polypeptide subunits in multimeric proteins define their quaternary structure.
- Cells contain large supramolecular assemblies, sometimes called molecular machines, in which all the necessary participants in complex cellular processes (e.g., DNA, RNA, and protein synthesis; photosynthesis; signal transduction) are bound together.
- Proteins with similar amino acid sequences generally can be assumed to have similar three-dimensional structures and similar functions. There are also examples of polypeptide chains with dissimilar sequences folding into similar three-dimensional structures.

- ### 3.2 Protein Folding

## Planar Peptide Bonds Limit the Shapes into Which Proteins Can Fold

$$\begin{array}{c}
 \text{O} \\
 \parallel \\
 \text{P}_1 - \text{C} - \text{N} - \text{P}_2 \\
 | \\
 \text{H}
 \end{array}
 \longleftrightarrow
 \begin{array}{c}
 \text{O}^- \\
 \parallel \\
 \text{P}_1 - \text{C} = \text{N}^+ - \text{P}_2 \\
 | \\
 \text{H}
 \end{array}
 \text{ or }
 \begin{array}{c}
 \text{O}^- \\
 \parallel \\
 \text{P}_1 - \text{C} = \text{N}^+ - \text{H} \\
 | \\
 \text{P}_2
 \end{array}$$

trans
cis

shapes—is rotation of the fixed planes of adjacent peptide bonds with respect to one another about two bonds: the C $\alpha$ -amino nitrogen bond (rotational angle called  $\Phi$ ) and the C $\alpha$ -carbonyl carbon bond (rotational angle called  $\Psi$ ).

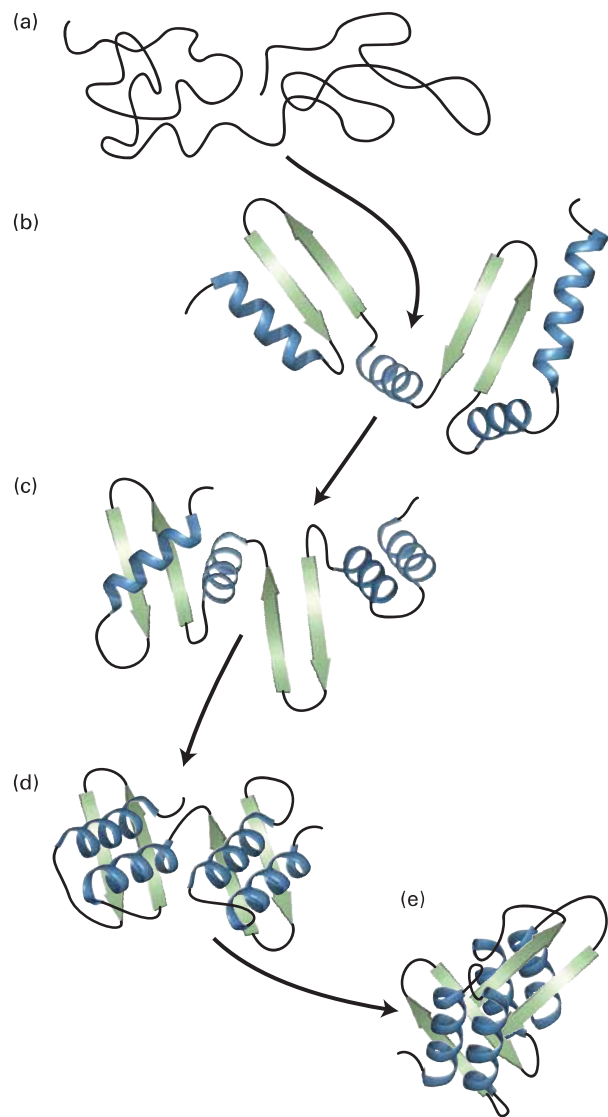
## The Amino Acid Sequence of a Protein Determines How It Will Fold

What features of natively well-ordered proteins limit their folding from so many potential conformations to just one or a few? The properties of the side chains (e.g., size, hydrophobicity, ability to form hydrogen and ionic bonds), together with their particular sequence along the polypeptide backbone, impose key restrictions. For example, a large side chain, such as that of tryptophan, might sterically block

one region of the chain from packing closely against another region, whereas a side chain with a positive charge, such as that of arginine, might attract a segment of the polypeptide that has a complementary negatively charged side chain (e.g., aspartic acid). Another example we have already discussed is the effect of the aliphatic side chains in heptad repeats in promoting the association of helices and the consequent formation of coiled coils. Thus a polypeptide's primary structure determines its secondary, tertiary, and quaternary structures.

The initial evidence that the information necessary for a protein to fold properly is encoded in its amino acid sequence came from *in vitro* studies (in test tubes) on the refolding of purified proteins, especially the Nobel Prize-winning studies in the 1960s by Christian Anfinsen of the refolding of ribonuclease A, an enzyme that cleaves RNA. Others had previously shown that various chemical and physical perturbations can disrupt the weak noncovalent interactions that stabilize the native conformation of a protein, leading to the loss of its normal tertiary structure. The disruption of a protein's structure (and this can include secondary as well as tertiary structure) is called **denaturation**. Denaturation can be induced by thermal energy from heat, extremes of pH that alter the charges on amino acid side chains, or exposure to *denaturants* such as urea or guanidine hydrochloride at concentrations of 6–8 M, all of which disrupt structure-stabilizing noncovalent interactions. Treatment with reducing agents, such as  $\beta$ -mercaptoethanol, that break disulfide bonds can further destabilize disulfide-containing proteins. Under denaturing conditions, a population of uniformly folded protein molecules is destabilized and converted into a collection of many unfolded, or denatured, molecules that have many different non-native and biologically inactive conformations. As we have seen, a large number of possible non-native conformations exist (e.g.,  $8^n - 1$ ). There are two broad classes of non-native conformations seen in proteins: (1) monomeric unfolded or denatured structures and (2) aggregates, which can either be amorphous or have a well-organized structure, as is the case for the disease-associated amyloid fibrils described later in this chapter. In principle, aggregates can comprise many copies of a single protein (homogeneous aggregates) or contain a mixture of distinct proteins (heterogeneous aggregates).

The spontaneous unfolding of proteins under denaturing conditions is not surprising, given the substantial increase in entropy that occurs because a denatured protein can adopt many non-native conformations (increased disorder). What is striking, however, is that when a pure sample of a single type of unfolded protein in a test tube is shifted back very carefully to normal conditions (body temperature, normal pH levels, reduction in the concentration of denaturants), some denatured polypeptides can spontaneously refold into their native, biologically active states, as in Anfinsen's experiments. This kind of refolding experiment, as well as studies showing that synthetic proteins made chemically can fold properly, established that the information contained in a protein's primary structure can be sufficient to direct correct refolding. Newly synthesized proteins appear to fold into their proper conformations just as denatured proteins



**FIGURE 3-16 Hypothetical protein-folding pathway.** Folding of a monomeric protein follows the structural hierarchy of primary (a) → secondary (b–d) → tertiary (e) structure. Formation of small structural motifs (c) appears to precede formation of domains (d) and the final tertiary structure (e).

do. The observed similarity in the folded, three-dimensional structures of proteins with similar amino acid sequences, noted in Section 3.1, provided additional evidence that the primary sequence also determines protein folding *in vivo* (in live organisms). It appears that formation of secondary structures and structural motifs occurs early in the folding process, followed by assembly of more complex structural domains, which then associate into more complex tertiary and quaternary structures (Figure 3-16).

### Folding of Proteins in Vivo Is Promoted by Chaperones

The conditions under which a purified, denatured protein refolds in a test tube differ markedly from the conditions



under which a newly synthesized polypeptide folds in a cell. The presence of other biomolecules, some of which are themselves nascent and in the process of folding, can potentially interfere with the autonomous, spontaneous folding of an otherwise natively well-ordered protein by forming aggregates. The cytosolic concentrations of some proteins are very high, and the total cytosolic protein concentration can be ~300 mg/ml in mammalian cells. These high protein concentrations favor aggregate formation by increasing the chances a nascent protein will encounter other proteins prior to completing its folding. Unfolded and partly folded proteins tend to aggregate into large, often water-insoluble masses, from which it is extremely difficult for a protein to dissociate and then fold into its proper conformation. In part, this aggregation is due to the exposure of hydrophobic side chains that have not yet had a chance to be buried in the inner core of the folded protein. Exposed hydrophobic side chains on different molecules will stick to one another, owing to the hydrophobic effect (see Chapter 2), and thus promote aggregation. The risk of such aggregation is especially high for newly synthesized proteins that have not yet completed their proper folding. Intrinsically disordered proteins are much less likely to form aggregates because, at least in some cases, they have relatively fewer hydrophobic side chains that can mediate such aggregation. Although protein folding into a well-ordered native state can occur *in vitro*, this does not happen for all unfolded molecules in a timely fashion because of the very large number of potentially incorrect, intermediate conformations into which the protein might fold.

Given such impediments, cells require faster, more efficient mechanisms for folding natively well-ordered proteins into their correct shapes than sequence alone provides. Without such help, cells might waste much energy in the synthesis of improperly folded, nonfunctional proteins, which would have to be destroyed to prevent their disrupting cell function. Cells clearly have such mechanisms, since more than 95 percent of the proteins present within cells have been shown to be in their native conformations. Proteins that do not or cannot fold properly—for example, those encoded by genes with mutations that alter the amino acid sequence—are often recognized as unfolded and rapidly degraded (hydrolyzed) by enzymes. The explanation for the cell's remarkable efficiency in promoting proper protein folding is that cells make a set of proteins, called **chaperones**, that facilitate proper folding of nascent proteins. One way chaperones facilitate proper folding is to prevent aggregation by binding to the target polypeptide or sequestering it from other partially or fully unfolded proteins, thus giving the nascent protein time to fold properly. The importance of chaperones is highlighted by the observation that many are evolutionarily conserved. Chaperones are found in all organisms from bacteria to humans, and some are homologs with high sequence similarity that use almost identical mechanisms to assist protein folding.

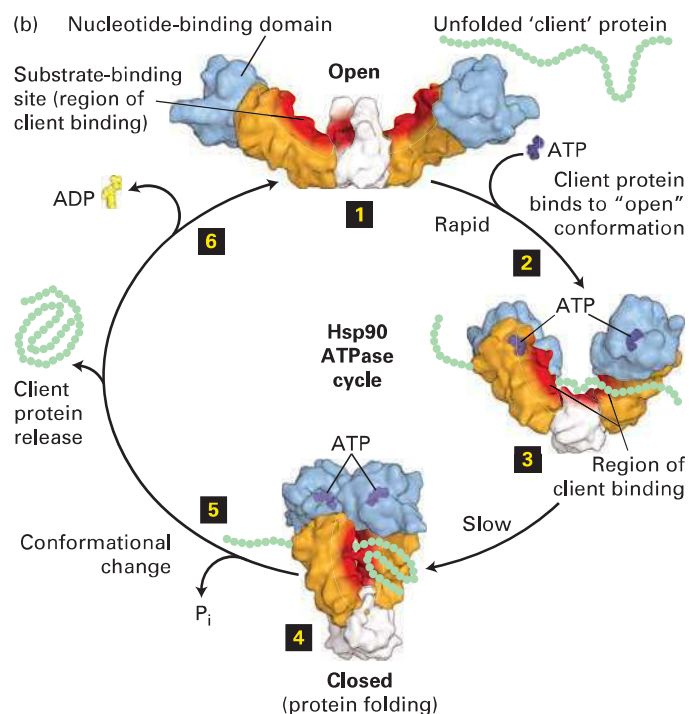
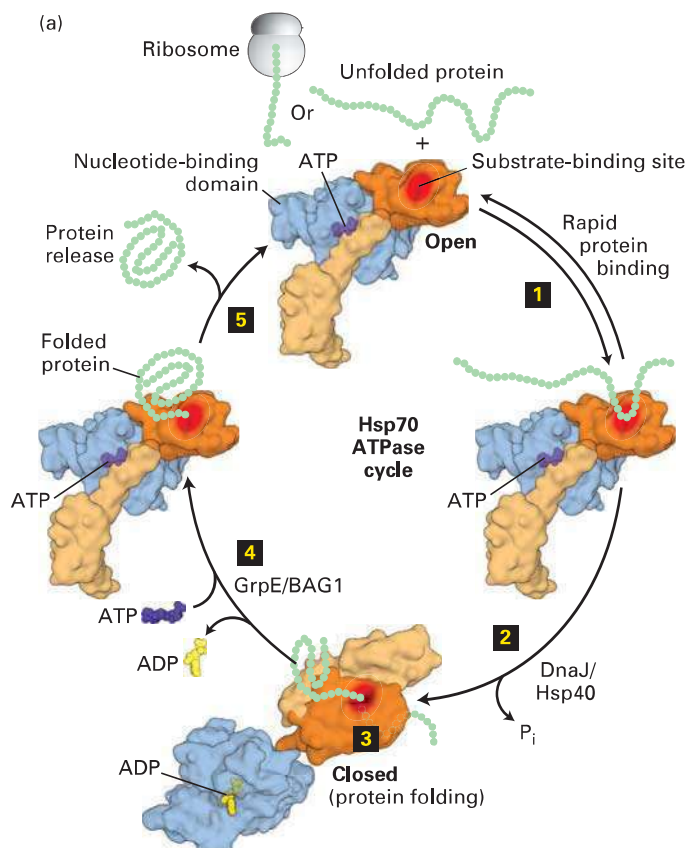
Chaperones can fold newly made proteins into functional conformations, refold misfolded or unfolded proteins into functional conformations, disassemble potentially toxic protein aggregates that form due to protein misfolding,

assemble and dismantle large multiprotein complexes, and mediate transformations between inactive and active forms of some proteins. Chaperones, which in eukaryotes are located in every cellular compartment and organelle, bind to the target proteins—also called substrates or client proteins—whose folding they will assist. Chaperones use a cycle of ATP binding, ATP hydrolysis to ADP, and exchange of a new ATP molecule for the ADP to induce a series of conformational changes that are essential for their function. There are several different classes of chaperones with distinct structures, all of which use ATP binding and hydrolysis in a variety of ways, which include (1) enhancing the binding of the target protein and (2) switching their own conformation. This ATP-dependent conformational switching is used (1) to optimize folding, (2) to return the chaperone to its initial state so that it is available to help fold another molecule, and (3) to set the time permitted for refolding, which can be determined by the rate of ATP hydrolysis.

Two general families of chaperones have been identified:

- **Molecular chaperones**, which bind to a short segment of a protein substrate and stabilize unfolded or partly folded proteins, thereby preventing these proteins from aggregating and being degraded.
- **Chaperonins**, which form folding chambers into which all or part of an unfolded protein can be sequestered, giving it time and an appropriate environment to fold properly.

**Molecular Chaperones** The heat-shock protein Hsp70 in the cytosol and its homologs (Hsp70 in the mitochondrial matrix, BiP in the endoplasmic reticulum, and DnaK in bacteria) are molecular chaperones. They were first identified by their rapid appearance after a cell had been stressed by heat shock (*Hsp* stands for “heat-shock protein”). Hsp70 and its homologs are the major chaperones in all organisms that use an ATP-dependent cycle to fold their substrates (Figure 3-17a). When bound to ATP, the monomeric Hsp70 protein assumes an open conformation, in which an exposed hydrophobic substrate-binding pocket transiently binds to exposed hydrophobic regions of an incompletely folded or partially denatured target protein, and then rapidly releases this substrate, as long as ATP is bound (step 1 in Figure 3-17a). Hydrolysis of the bound ATP causes the molecular chaperone to assume a closed form that binds its substrate protein much more tightly, and this tighter binding appears to facilitate the target protein's folding, in part by preventing it from aggregating with other unfolded proteins (step 2 in Figure 3-17a). Next the exchange of ATP for the chaperone-bound ADP (step 3) causes a conformational change in the chaperone that releases the target protein and regenerates an “empty,” ATP-bound Hsp70 ready to help fold another protein (step 4). If the target is now properly folded, it cannot rebind to an Hsp70. If it remains at least partially unfolded, it can bind again to give a chaperone another chance to help fold it properly. As we will see later in this chapter, a variety of proteins use a cycle of trinucleotide hydrolysis to a dinucleotide, followed by



**FIGURE 3-17 Molecular chaperone-mediated protein folding.**

(a) Hsp70. Many proteins fold into their proper three-dimensional structures with the assistance of Hsp70 or one of several Hsp70-like proteins. These molecular chaperones transiently bind to a nascent polypeptide as it emerges from a ribosome or to a protein that has otherwise unfolded. In the Hsp70 cycle, an unfolded substrate protein binds in rapid equilibrium (step **1**) to Hsp70's substrate-binding site (red) in the open conformation of its substrate-binding domain (light and dark orange) when an ATP (purple) is bound at Hsp70's nucleotide-binding domain (light blue). The substrate-binding domain comprises two subdomains (light and dark orange) that change relative positions and conformations during the cycle. Co-chaperone accessory proteins (DnaJ/Hsp40) stimulate the hydrolysis of ATP to ADP (yellow) that induces a large conformational change in the substrate-binding domain, resulting in the closed conformation, in which the substrate is locked into the substrate-binding domain; here proper folding is facilitated (steps **2** and **3**). Exchange of ATP for the bound ADP, stimulated by other accessory co-chaperone proteins (GrpE/BAG1), converts the Hsp70 back to the open conformation (step **4**), releasing the properly folded substrate (step **5**) and regenerating the open conformation, which can then interact with additional substrates. (b) Three conformational states of the dimeric Hsp90 molecular chaperone thought to be involved in substrate (also called client) remodeling. Client proteins bind at the substrate-binding

site (red surface) shared by the substrate-binding (orange) and C-terminal dimerization (white) domains and are thought to be remodeled in response to ATP binding and hydrolysis. The Hsp90 cycle begins when there is no nucleotide bound to the nucleotide-binding domains (light blue) and the dimer is in a very flexible, open configuration (step **1**) that can bind a client. Rapid ATP binding leads to a conformational change (step **2**) in which the nucleotide-binding domains and the substrate-binding domains move together (intermediate shown in step **3**) into a closed conformation in which the nucleotide-binding domains are dimerized (step **4**). The precise locations in Hsp90 at which clients bind apparently vary for different clients, but the binding surface, including the intersection of the substrate-binding domains and C-terminal dimerization domains (highlighted by red shading) binds a number of clients. ATP hydrolysis results in a conformational change in Hsp90 (step **5**) that may include a highly compact form, folding of the client, and client protein release. The ADP-bound form of Hsp90 can adopt several conformations, including a highly compact form. Release of ADP (step **6**) regenerates the initial flexible open state, which can then interact with additional clients. See E. D. Kirschke et al., 2014, *Cell* **157**:1685 and M. Taipale, D. F. Jarosz, and S. Lindquist, 2010, *Nat. Rev. Mol. Cell Biol.* **11**:515. [Solvent-accessible surface model of HSP90 courtesy of Elaine Kirschke and David A. Agard, UCSF. Open (ATP) PDB ID 2ior, closed (ATP) PDB ID 2cg9, closed (ADP) based on PDB ID 2cg9.]

dinucleotide/trinucleotide exchange, to control their activities. Later in this chapter, we will discuss a group of proteins called GTPases that depend on the exchange of GTP, rather than ATP, for bound GDP (instead of ADP) to induce conformational changes that dramatically influence the proteins' activities and the subsequent hydrolysis of the bound GTP to GDP.

Additional proteins, such as the co-chaperone Hsp40 in eukaryotes (DnaJ in bacteria), help increase the efficiency of the Hsp70-mediated folding of many proteins not only by stimulating the binding of substrate, but also by increasing the rate of hydrolysis of ATP by 100- to 1000-fold (see step **2** in Figure 3-17a). Members of four different families of nucleotide

exchange factors (e.g., GrpE in bacteria; BAG, HspBP, and Hsp110 in eukaryotes) also interact with Hsp70 (or DnaK), promoting the exchange of ATP for ADP (see step **3**). Multiple molecular chaperones are thought to bind to all nascent polypeptide chains as they are being synthesized on ribosomes. In bacteria, 85 percent of the proteins are released from their chaperones and proceed to fold normally; an even higher percentage of proteins in eukaryotes follow this pathway.

The Hsp70 protein family is not the only class of molecular chaperones. Another distinct class of molecular chaperones is the Hsp90 family, whose members usually recognize partially folded substrate proteins. Evolutionarily related Hsp90 family members are present in all organisms except archaea. Their strong evolutionary conservation is seen in the high amino acid sequence similarity (55 percent) of the Hsp90 from the bacterium *E. coli* and human Hsp90. In most eukaryotes, there are four distinct Hsp90s, two of which are in the cytosol (at 1–2 percent of total protein, Hsp90 is one of the most abundant cytosolic proteins) and one each in the endoplasmic reticulum and the mitochondrion. Although the range of protein substrates for Hsp90 chaperones is not as broad as for some other chaperones (at least 10 percent of yeast proteins are thought to be Hsp90 substrates), the Hsp90s are essential in eukaryotes. The Hsp90s help cells cope with denatured proteins generated by stress (e.g., heat shock), and they ensure that some of their substrates, usually called “clients,” can be converted from an inactive to an active state or otherwise held in a functional conformation. In some cases, an Hsp90 forms a relatively stable complex with a client until an appropriate signal causes its dissociation from the client, freeing the client to perform some regulated function in the cell. Hsp90 clients include transcription factors such as the receptors for the steroid hormones estrogen and testosterone. These steroid receptors regulate sexual development and function by controlling the activities of many genes (see Chapter 9). Another type of Hsp90 client is the set of enzymes called kinases, which control the activities of many proteins by phosphorylation (see Chapters 15 and 16).

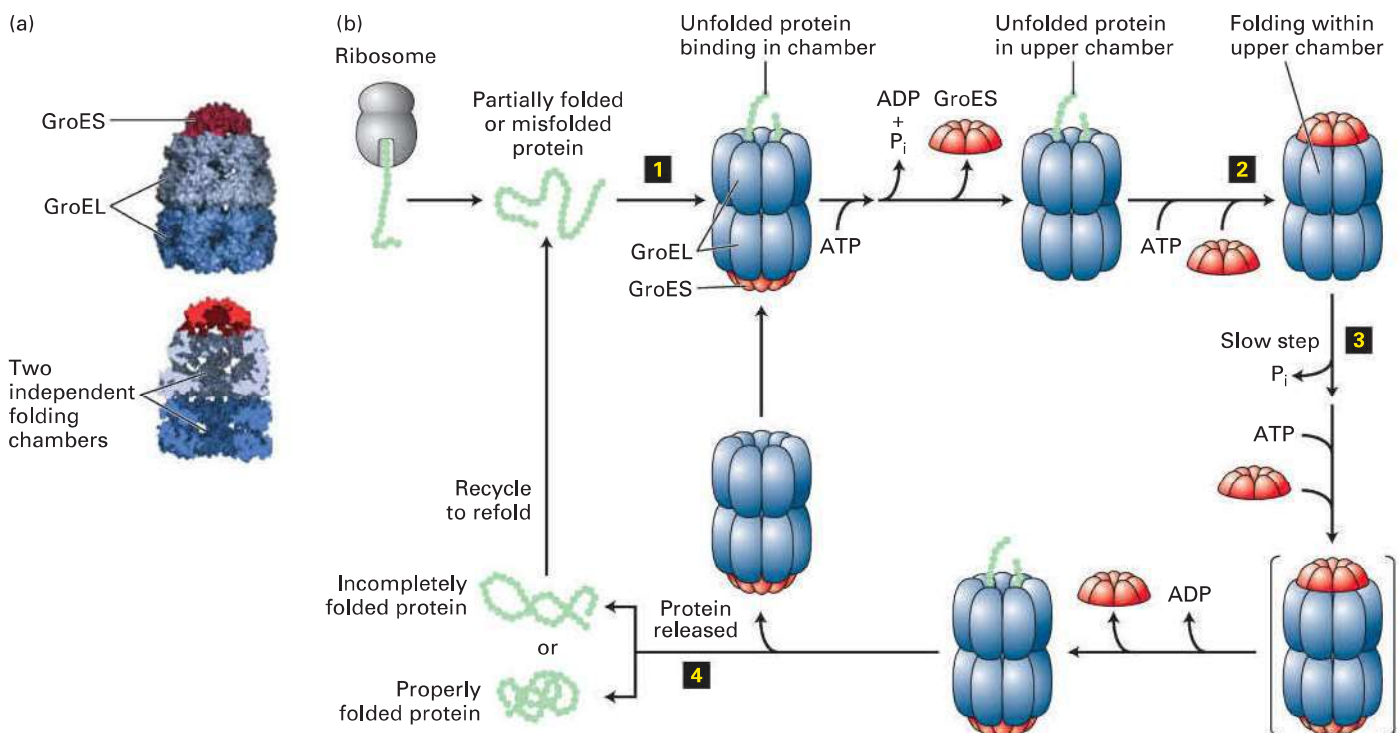
Unlike monomeric Hsp70, Hsp90 functions as a dimer in a cycle in which ATP binding, hydrolysis, and ADP release are coupled to major conformational changes and to binding, folding or activation, and release of clients (Figure 3-17b). Although much about the mechanism of Hsp90 remains to be learned, it is clear that clients bind to the substrate-binding domains when the chaperone is in the “open” conformation (step **1** in Figure 3-17b), that ATP binding leads to interaction of the ATP-binding domains and formation of a “closed” conformation (steps **1** and **2** in Figure 3-17b), and that hydrolysis of ATP plays an important role in activation of some client proteins and their subsequent release from the Hsp90 (step **3**). We also know that there are at least 20 co-chaperones that can have profound effects on the activity of Hsp90, including modulating its ATPase activity and determining which proteins will be clients (client specificity). Co-chaperones can also help coordinate the activities of Hsp90 and Hsp70. For example, Hsp70 can help begin the folding of a client that is then handed off by a co-chaperone to Hsp90 for additional

processing. Hsp90 activity can also be influenced by its covalent modification by small molecules. Finally, Hsp90 can help cells recognize misfolded proteins that are unable to refold and facilitate their degradation by mechanisms discussed later in this chapter. Thus, as part of the quality-control system in cells, molecular chaperones can help properly fold proteins or facilitate the destruction of those that cannot fold properly.

**Chaperonins** The proper folding of a large variety of newly synthesized proteins also requires the assistance of another class of proteins, the chaperonins, also called Hsp60s. These huge cylindrical supramolecular assemblies are formed from two rings of oligomers. There are two distinct groups of chaperonins that differ somewhat in their structures, detailed molecular mechanisms, and locations. Group I chaperonins, found in prokaryotes, chloroplasts, and mitochondria, are composed of two rings, each having seven subunits that interact with a homoheptameric co-chaperone “lid.” The bacteria group I chaperonin, known as GroEL/GroES, is shown in Figure 3-18a. In the bacterium *E. coli*, GroEL is thought to participate in the folding of about 10 percent of all proteins. Group II chaperonins, which are found in the cytosol of eukaryotic cells (e.g., TriC in mammals) and in archaea, can have eight to nine either homomeric or heteromeric subunits in each ring, and the “lid” function is incorporated into those subunits themselves—no separate lid protein is needed. It appears that ATP hydrolysis triggers the closing of the lid of group II chaperonins.

Figure 3-18b illustrates the GroEL/GroES cycle of protein folding. A partially folded or misfolded polypeptide of less than 60 kDa in mass is captured by hydrophobic residues near the entrance of the GroEL chamber and enters one of the folding chambers (upper chamber in Figure 3-18b). The second chamber is blocked by a GroES lid. Each of the 14 subunits of GroEL can bind ATP, hydrolyze it, and subsequently release ADP. These reactions are concerted for each set of seven subunits in a single ring and lead to major conformational changes. These changes control both the binding of the GroES lid that seals the chamber and the environment of the chamber in which polypeptide folding takes place. The polypeptide remains encased in the chamber capped by the lid. There it can undergo folding until ATP hydrolysis in that chamber, which is the slowest, rate-limiting step in the cycle ( $t_{1/2} \sim 10$  s), induces binding of ATP and a different GroES to the other ring. This then causes the GroES lid and ADP bound to the peptide-containing ring to be released, opening the chamber and permitting the folded protein to diffuse out of the chamber. If the polypeptide is folded properly, it can proceed to function in the cell. If it remains partially folded or misfolded, it can rebind to an unoccupied GroEL and the cycle can be repeated. There is a reciprocal relationship between the two rings in one GroEL complex. The capping of one chamber by GroES to permit sequestered substrate folding in that chamber is accompanied by the release of substrate polypeptide from the chamber of the second ring (simultaneous binding, folding, and release from the second chamber is not illustrated in Figure 3-18b). There is a striking similarity between the capped-barrel design of GroEL/GroES, in which proteins are sequestered for





**FIGURE 3-18 Chaperonin-mediated protein folding.** Proper folding of some proteins depends on chaperonins such as the prokaryotic group I chaperonin GroEL. (a) GroEL is a barrel-shaped complex of fourteen identical  $\sim 60,000$ -MW subunits, arranged in two stacked rings (blue) of seven subunits each that form two distinct internal polypeptide folding chambers. Homoheptameric lids (10,000-MW subunits), GroES (red), can bind to either end of the barrel and seal the chamber on that side. (b) The GroEL-GroES folding cycle. A partly folded or misfolded polypeptide enters one of the folding chambers (step **1**). The second chamber is blocked by a GroES lid. Each ring of seven GroEL subunits binds seven ATPs, hydrolyzes them, and then releases the ADPs in a set order coordinated with GroES binding and release and polypeptide binding, folding, and release. The major conformational changes that take place in the GroEL

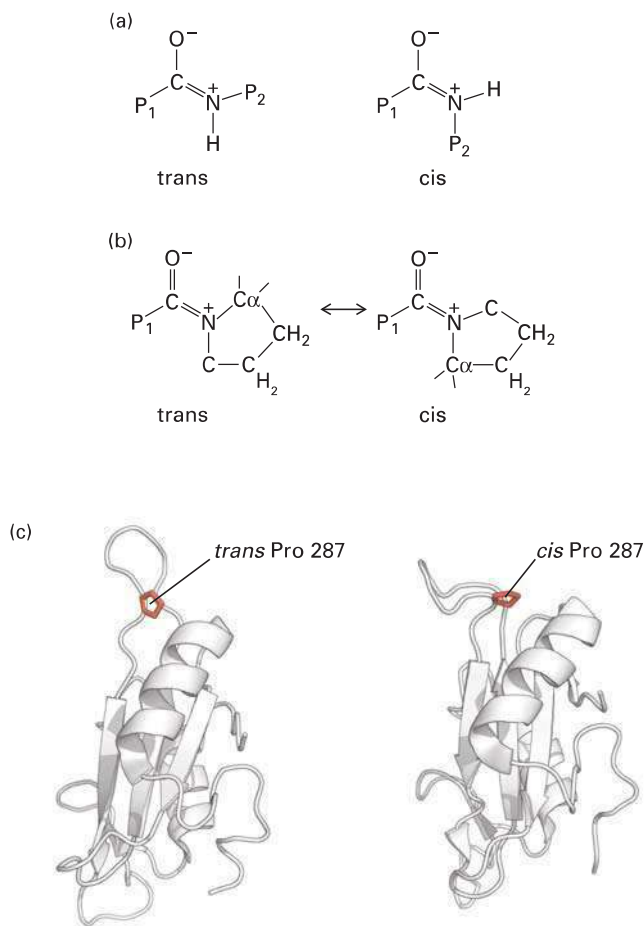
rings control the binding of the GroES lid that seals the chamber (step **2**). The polypeptide remains encased in the chamber capped by the lid, where it can undergo folding until ATP hydrolysis—the slowest, rate-limiting step in the cycle ( $t_{1/2} \sim 10$  s) (step **3**)—induces binding of ATP and a different GroES to the other ring (transient intermediate shown in brackets). This binding then causes the GroES lid and ADP bound to the peptide-containing ring to be released, opening the chamber and permitting the folded protein to diffuse out of the chamber (step **4**). If the polypeptide has folded properly, it can proceed to function in the cell. If it remains partially folded or misfolded, it can rebind to an unoccupied GroEL and the cycle can be repeated. See D. L. Nelson and M. M. Cox, 2013, *Lehninger Principles of Biochemistry*, 6th ed., Macmillan. [Part (a) data from Z. Xu, A. L. Horwich, and P. B. Sieglar, 1997, *Nature* **388**:741–750, PDB ID 1aon.]

folding, and the structure of the 26S proteasome that participates in protein degradation (discussed in Section 3.4). In addition, a group of proteins that are part of the AAA<sup>+</sup> family of ATPases are composed of hexameric rings with a central pore into which substrates can enter for folding or unfolding or in some cases proteolysis; examples of these will be discussed in Section 3.4 and in Chapter 13.

### Protein Folding Is Promoted by Proline Isomerases


As we learned earlier, the portions of the polypeptide chain on either side of a peptide bond ( $P_1$  and  $P_2$ ) are almost always oriented in a trans configuration (Figure 3-19a).

However, the trans configuration is not dramatically more energetically favorable than a cis configuration when there is a proline at  $P_2$  (Figure 3-19b). Among those folded proteins whose structures have been determined, about 5 percent of peptide bonds with proline at  $P_2$  exhibit the cis configuration, as compared with 0.03 percent of all other peptide bonds without proline at  $P_2$ . As the rate of isomerization between the cis and trans configurations is relatively slow, cells use proline isomerase proteins to catalyze these cis/trans isomerizations to facilitate the folding with the proper isomer. Prolyl isomerizations have been proposed to act as switches to alter the conformation, and thus the activity, of already stably folded proteins. Indeed, such isomerizations can substantially alter the structure of some proteins (Figure 3-19c).



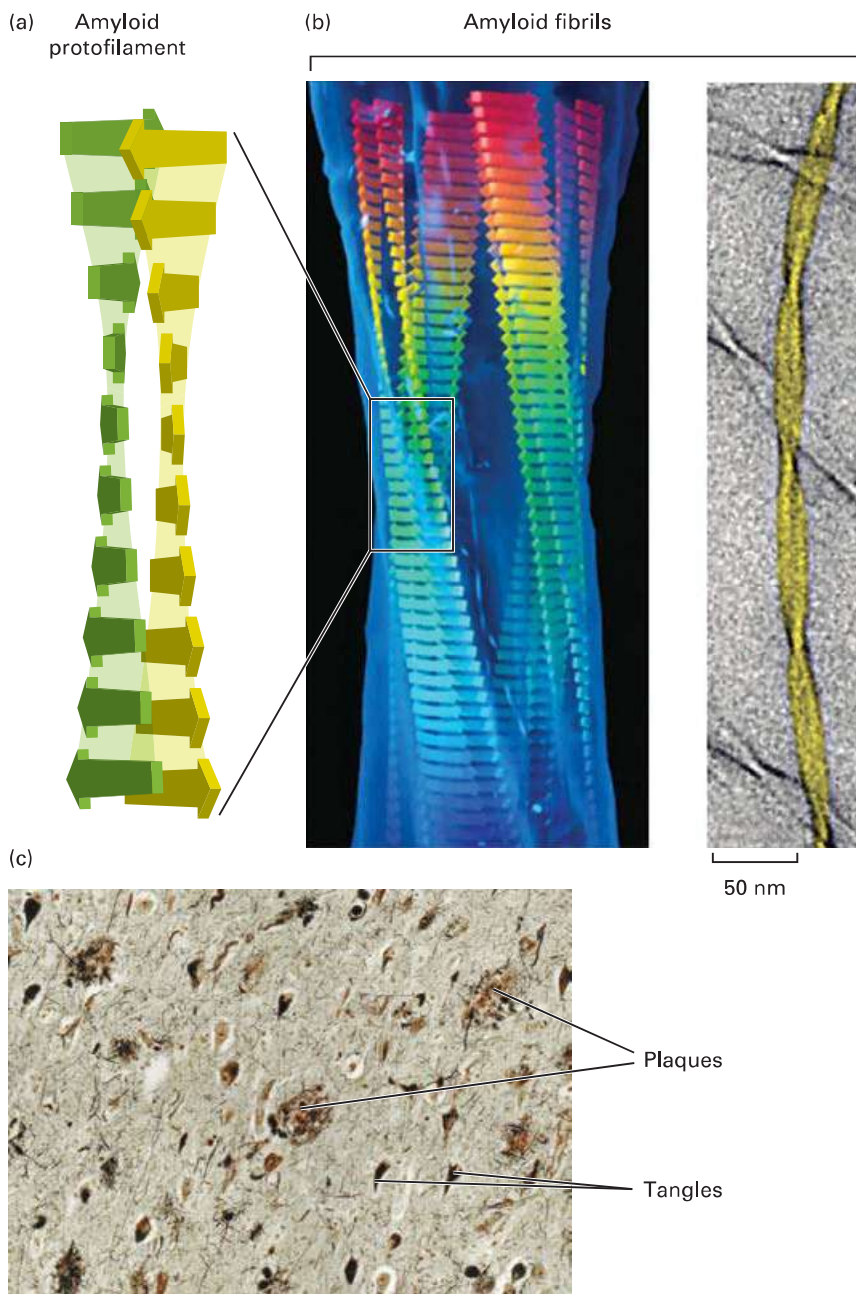
**FIGURE 3-19 Proline cis/trans isomerizations influence protein folding and structure.** (a) The planar, double bond-like character of peptide bonds leads to the potential of the portions of the polypeptide chain on either side (P<sub>1</sub> and P<sub>2</sub>) having cis or trans configurations. The trans configuration is present in about 99.97 percent of all peptide bonds in well-ordered proteins when P<sub>2</sub> is a residue other than proline. (b) When P<sub>2</sub> is proline, about 5 percent of peptide bonds are in the cis configuration. Proline isomerases catalyze the cis/trans isomerization to facilitate protein folding. (c) The structure of a portion of a protein, here an SH2 protein domain (see Chapter 16), can be dramatically altered by the cis/trans isomerization of a single proline, and this structural change can influence the protein's activity. [Part (c) trans data from E. V. Pletneva et al., 2006, *J. Mol. Biol.* **357**:550–561, PDB ID 2etz. Part (c) cis data from R. J. Mallis et al., 2002, *Nat. Struct. Biol.* **9**:900–905. PDB ID 1lui.]

### Abnormally Folded Proteins Can Form Amyloids That Are Implicated in Diseases

 After it is synthesized, a protein may fold into an alternative, abnormal three-dimensional structure as the result of mutations, inappropriate covalent modifications, or chemical (e.g., pH) or physical (e.g., heat)

alterations in its environment. Misfolding or denaturation can lead to a loss of the normal function of the protein and can result in the protein being marked for destruction (proteolytic degradation), as described later in this chapter. However, when degradation is incomplete or fails to keep pace with the production of misfolded protein, the misfolded protein or its proteolytic fragments can accumulate either inside or outside of cells in aggregates, or *plaques*, in various organs, including joints between bones, the liver, and the brain. Even those proteins or protein fragments that are normally highly resistant to aggregation, as is the case for intrinsically disordered proteins or protein fragments, will form aggregates if their concentrations are sufficiently elevated or when there are changes in environmental conditions. As noted above, such aggregates can either be amorphous or have a well-organized structure, which most commonly is the *amyloid state*. Strikingly, many diverse proteins can each aggregate into amyloid fibrils that have a common structure, called a cross- $\beta$  sheet (Figure 3-20a). Short segments, generally 6–12 residues long, in the unfolded or misfolded proteins hydrogen-bond to each other, forming a long array, or filament, of  $\beta$  sheets. In these arrays, each  $\beta$  strand is nearly perpendicular to the long axis of the filament, and two long, nearly flat  $\beta$  sheets pack closely together and twist around each other to form protofilaments, which then assemble together into thicker filaments, called *amyloid fibrils*. Within each protofilament the  $\beta$  strands can be either parallel or antiparallel (see Figure 3-5). Although some proteins form amyloid fibrils in their native, functional states, most amyloids are considered to be consequences of protein misfolding.

Amyloids were first recognized in protein aggregates that are deposited in tissues, are resistant to enzymatic degradation, and are associated with dozens of diseases, called amyloidoses. These diseases include neurodegenerative diseases such as Alzheimer's disease and Parkinson's disease in humans and transmissible spongiform encephalopathy ("mad cow" disease) in cows and sheep. Each of these diseases is characterized by the presence of filamentous plaques in a deteriorating brain (Figure 3-20b). Amyloidoses most commonly occur with aging; however, mutations in the genes encoding the aggregating protein can result in early amyloid formation and disease onset. The amyloid fibrils composing the plaques derive from abundant natural proteins. For example, fragments of the amyloid precursor protein, which is embedded in the plasma membrane, form the plaque found in the brains of patients with Alzheimer's disease; and prion protein, an "infectious" protein, forms fibrils in prion diseases. In Alzheimer's disease, a hyperphosphorylated form of the protein tau, normally a microtubule-binding protein (see Chapter 18), forms twisted fibers called "tangles." These amyloids, either as relatively short, water-soluble protofilaments or as long, insoluble fibrils, are thought to be toxic and to contribute directly to the pathology of amyloidoses. ■



**FIGURE 3-20 Misfolded proteins can form ordered amyloid aggregates based on a cross- $\beta$  sheet structure.** (a) In unfolded segments of proteins and polypeptides, exposed segments 6–12 residues long (short flat arrows) can assemble into  $\beta$  sheets (see also Figure 3-5) in which each  $\beta$  strand is oriented nearly perpendicularly to the long axis (vertical in this figure) of the resultant amyloid protofilament and hydrogen-bonded (light shading) to the strands above and below. Two long, nearly flat sheets pack closely together and twist around each other to form amyloid protofilaments, which then assemble together into thicker filaments called amyloid fibrils (b). Amyloid fibrils can be composed of varying numbers of protofilaments. A model of a four-protofilament-containing fibril fit into the electron density of acid-denatured insulin fibrils (left) and a cryoelectron microscopic image of two-protofilament-containing fibrils of fragments of transthyretin with an NMR-based model (yellow). Fibrils can aggregate into macroscopic plaques and tangles that are deposited in tissues and, when stained, are large enough to be visible using light microscopy. (c) Microscopic view of a section of human brain tissue from a patient with Alzheimer's disease with multiple amyloid plaques and fibrillary tangles. [Part (b, left) republished with permission of Elsevier, from Dobson, C.M., "Protein misfolding, evolution and disease," *Trends in Biochemical Science* 1999, **24**(9):329-332. Fig. 3. Part (b, right) reprinted by permission from Macmillan Publishers Ltd: from Knowles et al., *Nat. Rev. Mol. Cell Biol.* 2014, **15**(6):384-396. Fig. 3a. Part (c) Thomas Deerinck, NCMR/Science Source.]

## KEY CONCEPTS OF SECTION 3.2

### Protein Folding

- The primary structure (amino acid sequence) of a protein determines its three-dimensional structure, which determines its function. In short, function derives from structure; structure derives from sequence.
- Because protein function derives from protein structure, newly synthesized proteins must fold into the correct shape to function properly.
- The planar structure of the peptide bond limits the number of conformations a polypeptide can have (see Figure 3-15).

- The amino acid sequence of a protein dictates its folding into a specific three-dimensional conformation, the native state. Proteins will unfold, or denature, if treated under conditions that disrupt the noncovalent interactions stabilizing their three-dimensional structures.
- There are two broad classes of non-native conformations seen in misfolded or denatured proteins: (1) monomeric unfolded or denatured structures and (2) aggregates, which can either be amorphous or have a well-organized structure.
- Protein folding *in vivo* occurs with assistance from ATP-dependent chaperones. Chaperones can influence proteins in several ways, including preventing misfolding and



aggregation, facilitating proper folding, and maintaining an appropriate, stable structure required for subsequent protein activity (see Figure 3-17).

- There are two broad classes of chaperones: (1) molecular chaperones, which bind to a short segment of a substrate protein, and (2) chaperonins, which form folding chambers in which all or part of an unfolded protein can be sequestered, giving it time and an appropriate environment to fold properly. Cycles of ATP binding and hydrolysis, followed by exchange of the ADP produced with a new ATP molecule, play key roles in the mechanisms of protein folding by chaperones.
- Many misfolded or denatured proteins can form well-organized aggregates, called amyloid fibrils, made by short stretches of polypeptide that form a long array of  $\beta$  sheets nearly perpendicular to the fibril axis, called a cross- $\beta$  structure. Formation of amyloid fibrils that are resistant to degradation by diverse enzymes is associated with dozens of diseases called amyloidoses. Examples include the neurodegenerative diseases Alzheimer's disease and Parkinson's disease.

### 3.3 Protein Binding and Enzyme Catalysis

Proteins perform an extraordinarily diverse array of activities both inside and outside cells, yet most of these diverse functions are based on the ability of proteins to engage in a common activity: binding. Proteins bind to one another, to other macromolecules, to small molecules, and to ions. In this section, we describe some key features of protein binding and then turn to look at one group of proteins, enzymes, in greater detail. The activities of the other functional classes of proteins (structural, scaffold, transport, regulatory, motor) will be described in later chapters.

#### Specific Binding of Ligands Underlies the Functions of Most Proteins

The molecule to which a protein binds is called its **ligand**. In some cases, ligand binding causes a change in the shape of a protein. Such conformational changes are integral to the mechanism of action of many proteins and are important in regulating protein activity.

Two properties of a protein characterize how it binds ligands. *Specificity* refers to the ability of a protein to bind one molecule or a very small group of molecules in preference to all other molecules. *Affinity* refers to the tightness or strength of binding, usually expressed as the dissociation constant ( $K_d$ ). The  $K_d$  for a protein-ligand complex, which is the inverse of the equilibrium constant  $K_{eq}$  for the binding reaction, is the most common quantitative measure of affinity (see Chapter 2). The stronger the interaction between a protein and ligand, the lower the value of  $K_d$ . Both the specificity and the affinity of a protein for a ligand depend on the structure of the *ligand-binding site*. For high-affinity and highly specific interactions to take place, the shape and

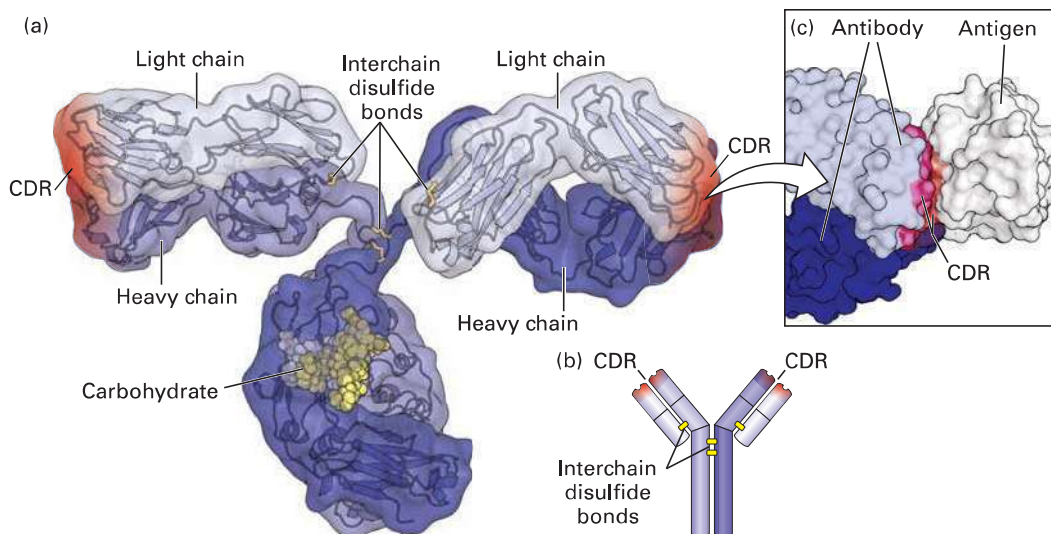
chemical properties of the binding site must be complementary to those of the ligand molecule, a property termed **molecular complementarity**. As we saw in Chapter 2, molecular complementarity allows molecules to form multiple noncovalent interactions at close range and thus stick together.

One of the best-studied examples of protein-ligand binding, involving high affinity and exquisite specificity, is the binding of antibodies to antigens. **Antibodies** are proteins that circulate in the blood and are made by the immune system in response to **antigens**, which are usually macromolecules present in infectious agents (e.g., a bacterium or a virus) or other foreign substances (e.g., proteins or polysaccharides in pollens). Different antibodies are generated in response to different antigens, and these antibodies have the remarkable characteristic of binding specifically to ("recognizing") the part of the antigen, called an **epitope**, that initially induced the production of the antibody, and not to other molecules. Antibodies act as specific sensors for antigens, forming antibody-antigen complexes that initiate a cascade of protective reactions in cells of the immune system. Chapter 23 discusses antibodies and their roles in the immune system, and later in this chapter we will discuss techniques for studying proteins that exploit antibodies. Here we briefly introduce the structure of antibodies and their binding to epitopes.

Antibodies are Y-shaped molecules, often formed from two identical longer, or *heavy*, chains and two identical shorter, or *light*, chains. In IgG antibodies (also called immunoglobulins, shown in Figure 3-21a), there are four globular domains in each heavy chain and two in each light chain, all of which are called immunoglobulin (Ig) domains. Each of the two branching arms of an IgG antibody contains a single light chain linked to a heavy chain by a disulfide bond, and two disulfide bonds covalently link the two heavy chains together. Near the end of each arm are six highly variable loops, called *complementarity-determining regions* (CDRs), which form the antigen-binding sites. The sequences of the six loops are highly variable among antibodies, generating unique complementary ligand-binding sites that make them specific for different epitopes (Figure 3-21b). The intimate contact between antibody and epitope surfaces, stabilized by numerous noncovalent interactions, is responsible for the extremely precise binding specificity exhibited by an antibody.

The specificity of antibodies is so precise that they can distinguish between the cells of individual members of a species and in some cases between proteins that differ by only a single amino acid, or even between proteins with identical sequences that differ only in their post-translational modifications. Because of their specificity and the ease with which they can be produced (see Chapter 23), antibodies are highly useful reagents used in many of the experiments discussed in subsequent chapters.

We will see many examples of protein-ligand binding throughout this book, including binding of hormones to receptors (see Chapter 15), binding of regulatory molecules to DNA (see Chapter 9), and binding of cell-adhesion molecules to extracellular matrices (see Chapter 20), to name just a few. Here we focus on how the binding of one class of proteins, enzymes, to their ligands results in the catalysis of the chemical reactions essential for the survival and function of cells.



**FIGURE 3-21 Protein-ligand binding of antibodies.** (a) Hybrid (surface and ribbon) model of an antibody. Every antibody molecule of the immunoglobulin G (IgG) class consists of two identical heavy chains (medium and dark blue) and two identical light chains (light blue) covalently linked by disulfide bonds (yellow). The complementarity-determining regions (CDRs) that define the antigen-binding sites are represented by red shading. (b) The cartoon shows the overall structure

containing the two heavy (longer) and two light (shorter) chains, with yellow bars representing disulfide bonds. (c) The hand-in-glove fit between an antibody and the site to which it binds (epitope) on its target antigen—in this case, chicken egg-white lysozyme. The antibody contacts the antigen with residues from its CDRs. [Part (a) data from L. J. Harris, et al., 1997, *Biochemistry* **36**:1581-1597, PDB ID 1igt. Part (b) data from E. A. Padlan et al., 1989, *P. Natl. Acad. Sci. USA* **86**:5938-5942, PDB ID 3hfm.]

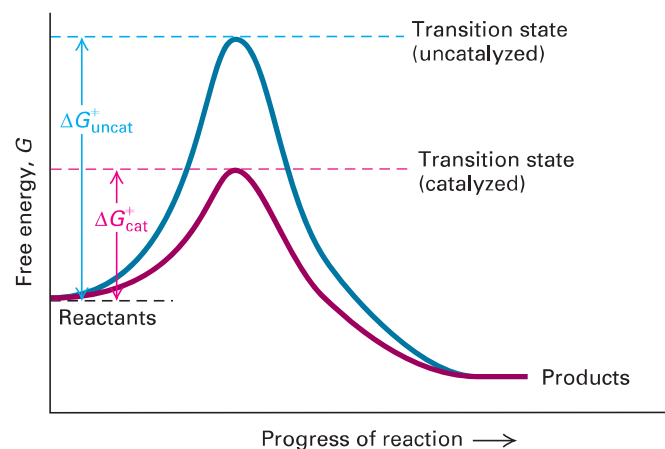
## Enzymes Are Highly Efficient and Specific Catalysts

Proteins that catalyze chemical reactions—the making and breaking of covalent bonds—are called **enzymes**, and the ligands of enzymes are called **substrates**. Enzymes make up a large and very important functional class of proteins—indeed, almost every chemical reaction in the cell is catalyzed by a specific catalyst, usually an enzyme. Another form of catalytic macromolecule in cells is made from RNA. These RNAs are called **ribozymes** (see Chapter 5).

Thousands of different types of enzymes, each of which catalyzes a single chemical reaction or a set of closely related reactions, have been identified. Certain enzymes are found in the majority of cells because they catalyze the synthesis of common cellular products (e.g., proteins, nucleic acids, and phospholipids) or take part in harvesting energy from nutrients (e.g., by the conversion of glucose and oxygen into carbon dioxide and water during cellular respiration). Other enzymes are present only in a particular type of cell because they catalyze chemical reactions unique to that cell type (e.g., the enzymes in neurons that convert tyrosine into dopamine, a neurotransmitter). Although most enzymes are located within cells, some are secreted and function at extracellular sites, such as the blood, the digestive tract, or even outside the organism (e.g., toxic enzymes in the venom of poisonous snakes).

Like all **catalysts** (see Chapter 2), enzymes increase the rate of a reaction, but they do not affect the extent of a reaction, which is determined by the change in free energy ( $\Delta G$ ) between reactants and products, and they are not themselves permanently changed as a consequence of the reaction they catalyze. Enzymes increase the reaction rate by lowering the energy of the *transition state*, and therefore the *activation energy*

required to reach it (Figure 3-22). In the test tube, catalysts such as charcoal and platinum facilitate reactions, but usually only at high temperatures or pressures, at extremes of high or low pH, or in organic solvents. Within cells, however, enzymes must function effectively in an aqueous environment at 37 °C



**FIGURE 3-22 Effect of an enzyme on the activation energy of a chemical reaction.** This hypothetical reaction pathway depicts the changes in free energy,  $G$ , as a reaction proceeds. A reaction will take place spontaneously only if the total  $G$  of the products is less than that of the reactants (negative  $\Delta G$ ). However, all chemical reactions proceed through one or more high-energy transition states, and the rate of a reaction is inversely proportional to the activation energy ( $\Delta G^\ddagger$ ), which is the difference in free energy between the reactants and the transition state (highest point along the pathway). Enzymes and other catalysts accelerate the rate of a reaction by reducing the free energy of the transition state and thus  $\Delta G^\ddagger$ .

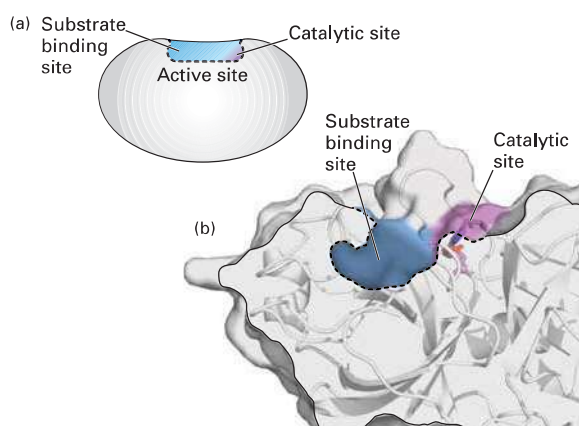
and 1 atmosphere of pressure and at physiological pH values, usually 6.5–7.5 but sometimes lower. Remarkably, enzymes exhibit immense catalytic power, in some cases accelerating the rates of reactions to  $10^6$ – $10^{12}$  times those of the corresponding uncatalyzed reactions under otherwise similar conditions.

## An Enzyme's Active Site Binds Substrates and Carries Out Catalysis

Certain amino acids of an enzyme are particularly important in determining its specificity and catalytic power. In the native conformation of an enzyme, critically important amino acids (which usually come from different parts of the linear sequence of the polypeptide) are brought into proximity, forming a cleft in the enzyme surface called the **active site** (Figure 3-23). An active site usually makes up only a small part of the total protein; the remaining part is involved in the folding of the polypeptide, regulation of the active site, and interactions with other molecules.

An active site consists of two functionally important regions: the *substrate-binding site*, which recognizes and binds the substrate or substrates, and the *catalytic site*, which carries out the chemical reaction once the substrate has bound. The catalytic groups in the catalytic site are amino acid side chains and backbone carboxyl and amino groups. In some enzymes, the catalytic and substrate-binding sites overlap; in others, the two regions are structurally distinct.

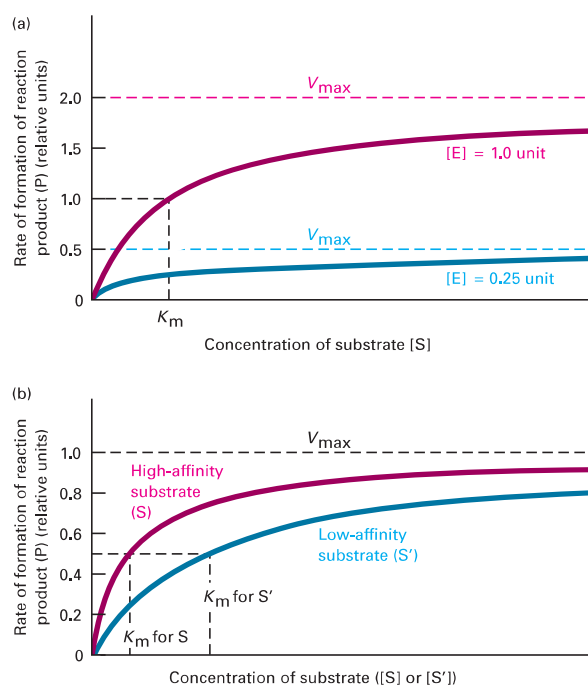
The substrate-binding site is responsible for the remarkable specificity of enzymes. Alteration of the structure of an enzyme's substrate by only one or a few atoms, or a subtle change in the geometry (e.g., stereochemistry) of the substrate, can result in a variant molecule that is no longer a substrate of the enzyme. As with the specificity of antibodies for antigens described above, the specificity of enzymes for substrates is a



**FIGURE 3-23 Active site of the enzyme trypsin.** (a) An enzyme's active site (outlined by dashed line) is composed of a substrate-binding site (blue), which binds specifically to a substrate, and a catalytic site (purple), which carries out catalysis. (b) A hybrid surface/ribbon representation of a portion of the serine protease trypsin. Clearly visible are the active-site cleft containing the catalytic site (purple, includes the key catalytic triad of Ser-195, Asp-102, and His-57, see also Figure 3-27) and a portion of the substrate-binding site called the side-chain-specificity binding pocket (blue). [Data from B. Sandler, M. Murakami, and J. Clardy, 1998, *J. Am. Chem. Soc.* **120**:595-596, PDB ID 1aq7.]

consequence of the precise molecular complementarity between an enzyme's substrate-binding site and the substrate. Usually only one or a few substrates can fit precisely into a binding site.

The idea that substrates might bind to enzymes in the manner of a key fitting into a lock was first suggested by Emil Fischer in 1894. A variation of this proposal by Daniel Koshland in 1958, called *induced fit*, posited that the substrate-binding site is not rigid, as a lock is, but flexible, and is induced to change shape for more optimal catalysis when the substrate binds. In 1913, Leonor Michaelis and Maud Leonora Menten provided crucial evidence supporting the enzyme-substrate binding hypothesis. They showed that the rate of an enzymatic reaction was proportional to the substrate concentration at low substrate concentrations, but that as substrate concentrations increased, the rate reached a plateau, or **maximal velocity**,  $V_{\max}$ , and became substrate concentration independent, with a value of  $V_{\max}$  directly proportional to the amount of enzyme present in the reaction mixture (Figure 3-24).



**FIGURE 3-24  $K_m$  and  $V_{\max}$  for an enzyme-catalyzed reaction.**  $K_m$  and  $V_{\max}$  are determined from analysis of the dependence of the initial reaction rate on substrate concentration. The shape of these hypothetical kinetic curves is characteristic of a simple enzyme-catalyzed reaction in which one substrate (S) is converted into product (P). The initial reaction velocity is measured immediately after addition of enzyme to substrate, before the substrate concentration changes appreciably. (a) Plots of initial reaction velocity at two different concentrations of enzyme [E] as a function of substrate concentration [S]. The [S] that yields a half-maximal reaction rate is the Michaelis constant  $K_m$ , a measure of the affinity of E for turning S into P. Quadrupling the enzyme concentration causes a proportional increase in the reaction rate, so the maximal velocity  $V_{\max}$  is quadrupled;  $K_m$ , however, is unaltered. (b) Plots of initial reaction velocity versus substrate concentration with a substrate S for which the enzyme has a high affinity and with a substrate S' for which the enzyme has a lower affinity. Note that  $V_{\max}$  is the same with both substrates because [E] is the same, but that  $K_m$  is higher for S', the low-affinity substrate.

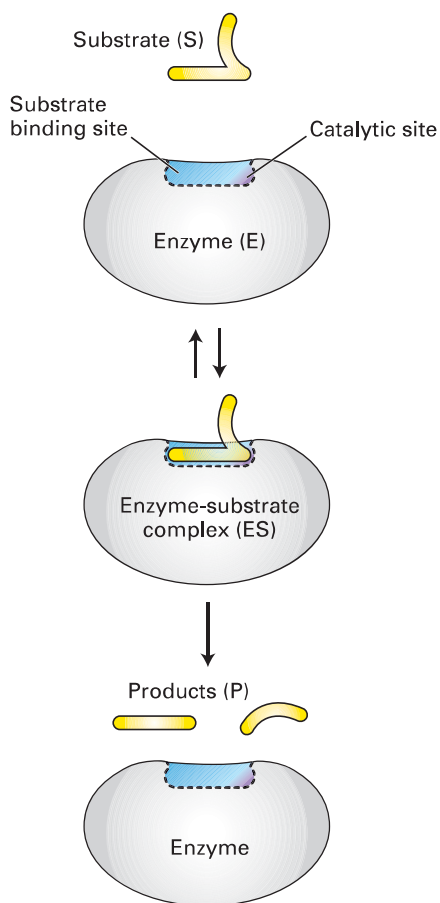


Michaelis and Menten deduced that these characteristics were due to the binding of substrate molecules (S) to a fixed and limited number of sites on the enzymes (E), and they called the bound species the enzyme-substrate (ES) complex. At high concentrations of substrate, all the binding sites on the enzymes have substrate bound, and the substrate-binding sites are said to be *saturated* with substrate—no additional binding to active sites is possible, and the maximal velocity of the reaction is achieved. Michaelis and Menten proposed that the ES complex is in equilibrium with the unbound enzyme and substrate and is an intermediate step in the ultimately irreversible conversion of substrate to product (P) (Figure 3-25):



and that the rate  $V_0$  of formation of product at a particular substrate concentration  $[S]$  is given by what is now called the *Michaelis-Menten equation*:

$$V_0 = V_{\max} \frac{[S]}{[S] + K_m} \quad (3-1)$$



**FIGURE 3-25 Schematic model of an enzyme's reaction mechanism.** Enzyme kinetics suggest that enzymes (E) bind substrate molecules (S) at a fixed and limited number of sites—the enzymes' active sites. The bound species is known as an enzyme-substrate (ES) complex. The ES complex is in equilibrium with the unbound enzyme and substrate (double arrows) and is an intermediate step in the conversion of substrate to products (P).

where the **Michaelis constant**,  $K_m$ , a measure of the affinity of an enzyme for its substrate, is the substrate concentration that yields a half-maximal reaction rate (i.e.,  $\frac{1}{2} V_{\max}$  in Figure 3-24). The  $K_m$  is somewhat similar in nature, but not identical, to the dissociation constant,  $K_d$  (see Chapter 2). The smaller the value of  $K_m$ , the more effective the enzyme is at making product from dilute solutions of substrate, and the lower the substrate concentration needed to reach half-maximal velocity. The smaller the  $K_d$ , the lower the ligand concentration needed to reach 50 percent of binding. The concentrations of the various small molecules in a cell vary widely, as do the  $K_m$  values for the different enzymes that act on them. A good rule of thumb is that the intracellular concentration of a substrate is often approximately the same as, or somewhat greater than, the  $K_m$  value of the enzyme to which it binds.

The rates of reaction at substrate saturation vary enormously among enzymes. The maximum number of substrate molecules converted to product at a single enzyme active site per second, called the *turnover number*, can be less than 1 for very slow enzymes. The turnover number for carbonic anhydrase, one of the fastest enzymes, is  $6 \times 10^5$  molecules per second.

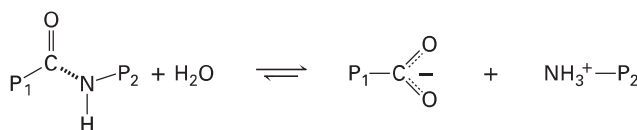
Many enzymes catalyze the conversion of substrates to products by dividing the process into multiple, discrete chemical reactions, in which the product of one reaction is the substrate for the subsequent reaction. These sequential reactions generate multiple, distinct enzyme-substrate complexes (ES, ES', ES'', etc.) prior to the final release of the products:

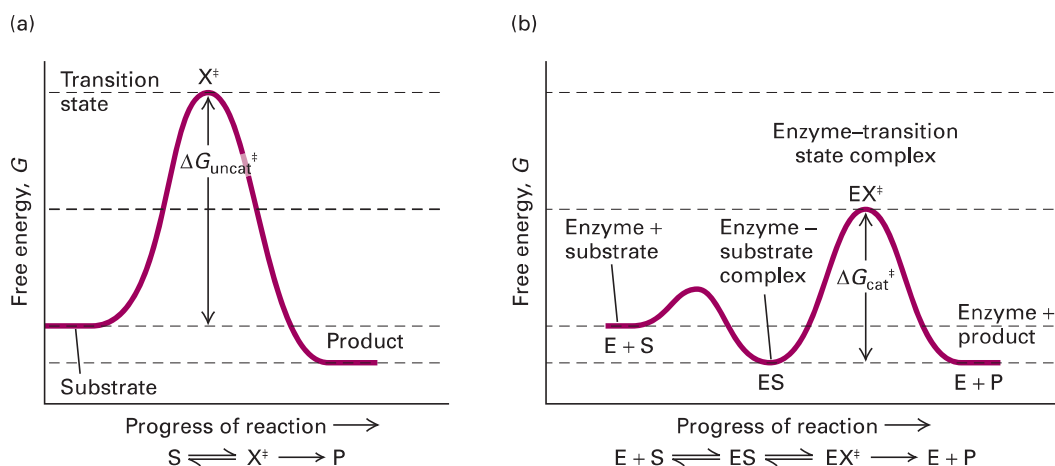


The energy profiles for such multistep reactions involve multiple hills and valleys (Figure 3-26). Methods have been developed to trap the intermediates in such reactions to learn more about the details of how enzymes catalyze reactions.

## Serine Proteases Demonstrate How an Enzyme's Active Site Works

Serine proteases, a large family of protein-cleaving, or proteolytic, enzymes, are used throughout the biological world—to digest meals (the pancreatic enzymes trypsin, chymotrypsin, and elastase), to control blood clotting (the enzyme thrombin), even to help silk moths chew their way out of their cocoons (cocoanase). This class of enzymes usefully illustrates how an enzyme's substrate-binding site and catalytic site cooperate in multistep reactions to convert substrate to product. Here we consider how trypsin and its two evolutionarily closely related pancreatic proteases, chymotrypsin and elastase, catalyze cleavage of a peptide bond in a polypeptide substrate:





**FIGURE 3-26 Free-energy reaction profiles of uncatalyzed and multistep enzyme-catalyzed reactions.** (a) The free-energy reaction profile of a hypothetical simple uncatalyzed reaction converting substrate (S) to product (P) via a single high-energy transition state. (b) Many enzymes catalyze such reactions by dividing the process

into multiple discrete steps, in this case, the initial formation of an ES complex followed by conversion via a single transition state ( $EX^\ddagger$ ) to the free enzyme (E) and P. The activation energy for each of these steps is significantly less than the activation energy for the uncatalyzed reaction; thus the enzyme dramatically enhances the reaction rate.

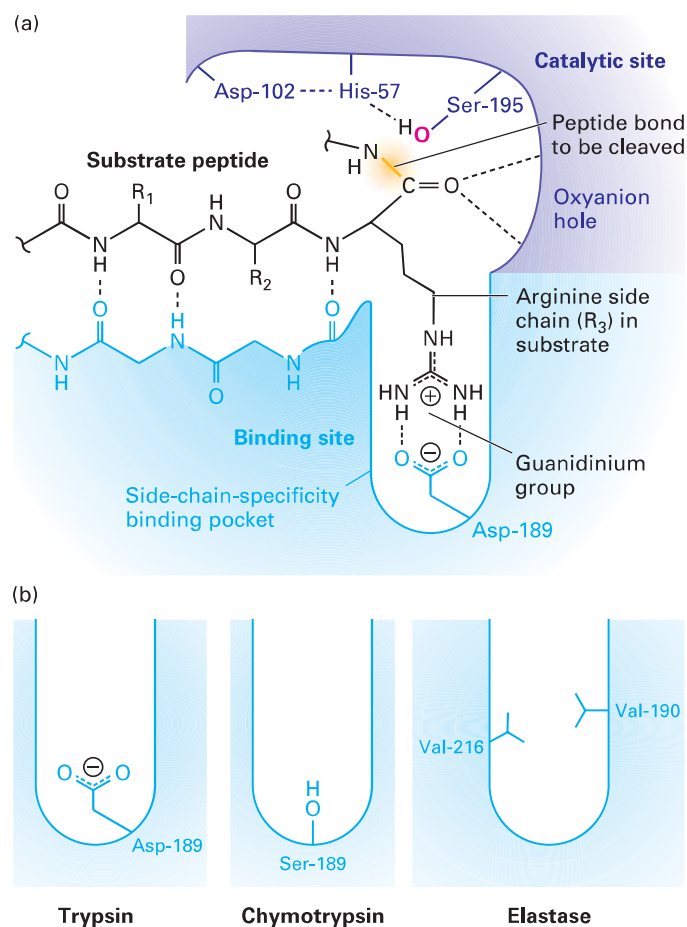
where  $P_1$  is the part of the protein on the N-terminal side of the peptide bond to be cleaved and  $P_2$  is the portion on the C-terminal side. We first consider how serine proteases bind specifically to their substrates and then show in detail how catalysis takes place.

Figure 3-27a shows how a substrate polypeptide binds to the substrate-binding site in the active site of trypsin. There are two key binding interactions. First, the substrate (black polypeptide backbone) and enzyme (blue polypeptide backbone) form hydrogen bonds that resemble those of a  $\beta$  sheet. Second, a key side chain of the substrate that determines which peptide in the substrate is to be cleaved extends into the enzyme's *side-chain-specificity binding pocket*, at the bottom of which resides the negatively charged side chain of trypsin's Asp-189. Trypsin has a marked preference for hydrolyzing substrates at the carboxyl ( $C=O$ ) side of an amino acid with a long, positively charged side chain (arginine or lysine) because the side chain is stabilized in the enzyme's side-chain-specificity binding pocket by the negative Asp-189.

**FIGURE 3-27 Substrate binding in the active site of trypsin-like serine proteases.** (a) The active site of trypsin (purple and blue molecule) with a bound substrate (black molecule). The substrate forms a two-stranded  $\beta$  sheet with trypsin's substrate-binding site, and the side chain of an arginine ( $R_3$ ) in the substrate is bound in the side-chain-specificity binding pocket of the binding site. Its positively charged guanidinium group is stabilized by the negative charge on the side chain of the enzyme's Asp-189. This binding aligns the peptide bond of the arginine appropriately for hydrolysis catalyzed by the enzyme's active-site catalytic triad (side chains of Ser-195, His-57, and Asp-102).

(b) The amino acids lining the side-chain-specificity binding pocket determine its shape and charge and thus its binding properties. Trypsin accommodates the positively charged side chains of arginine and lysine; chymotrypsin, large, hydrophobic side chains such as phenylalanine; and elastase, small side chains such as glycine and alanine. See J. J. Perona and C. S. Craik, 1997, *J. Biol. Chem.* **272**:29987–29990.

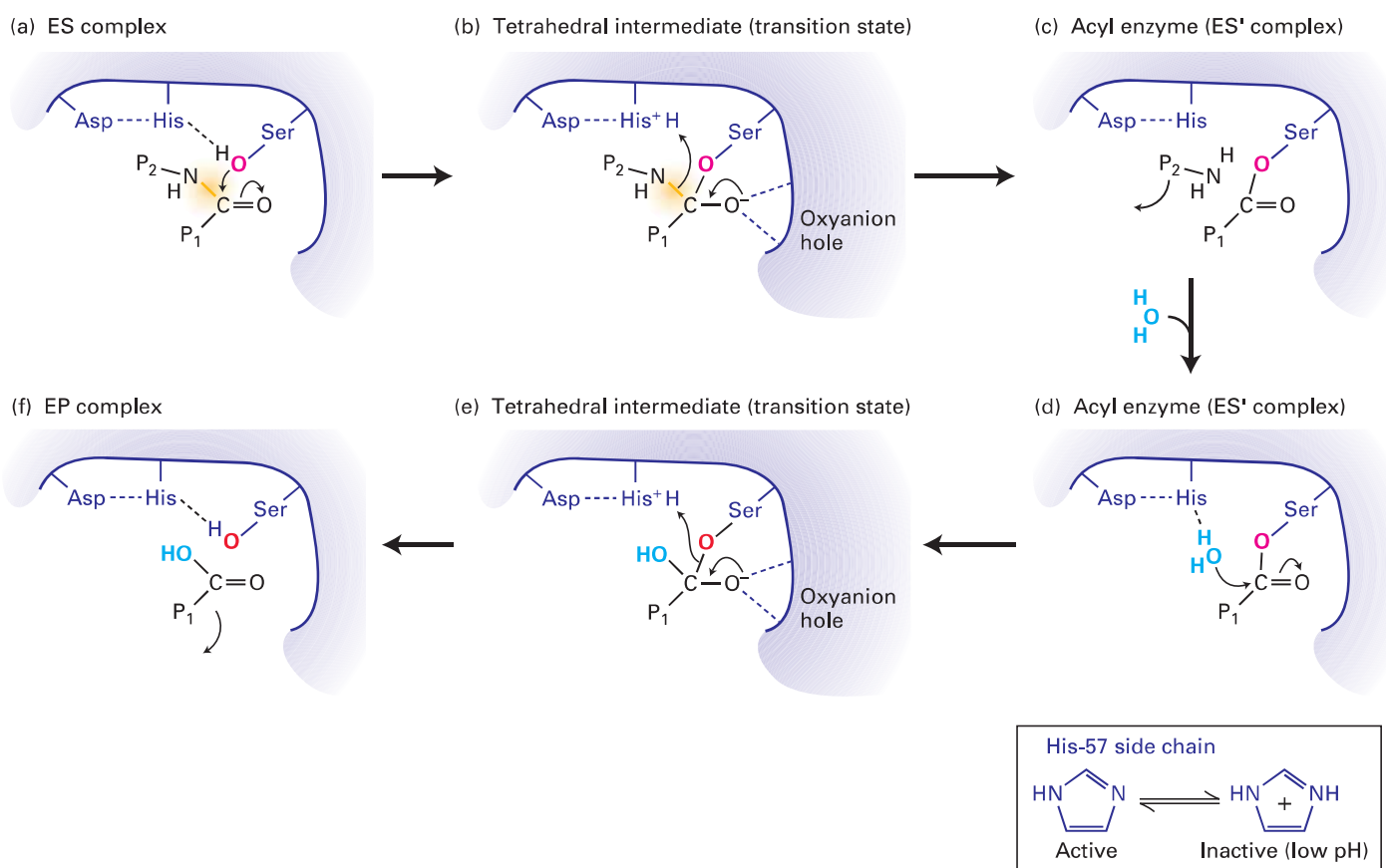
Slight differences in the structures of otherwise similar binding pockets help explain the differing substrate specificities of two serine proteases related to trypsin: chymotrypsin prefers large aromatic groups (as in Phe, Tyr, Trp), and elastase prefers the small side chains of Gly and Ala (Figure 3-27b). The uncharged Ser-189 in chymotrypsin



allows large, uncharged, hydrophobic side chains to bind stably in the binding pocket. The specificity of elastase is influenced by the replacement of glycines in the sides of the binding pocket in trypsin with the branched aliphatic side chains of valines (Val-216 and Val-190), which obstruct the binding pocket (see Figure 3-27b). As a consequence, large side chains in substrates are prevented from fitting into the binding pocket of elastase, whereas substrates with the short alanine or glycine side chains at this position can bind well and be subject to subsequent cleavage.

In the catalytic site, all three enzymes use the hydroxyl group on the side chain of a serine at position 195 to catalyze the hydrolysis of peptide bonds in substrate proteins. A catalytic triad formed by the three side chains of Ser-195, His-57, and Asp-102 participates in what is essentially a two-step hydrolysis reaction. Figure 3-28 shows how the

catalytic triad cooperates in breaking the peptide bond, with Asp-102 and His-57 supporting the attack of the hydroxyl oxygen of Ser-195 on the carbonyl carbon in the substrate. This attack initially forms an unstable transition state with four groups attached to the carbon (tetrahedral intermediate). Breaking of the C—N peptide bond then releases one part of the substrate protein ( $\text{NH}_3\text{—P}_2$ ), while the other part remains covalently attached to the enzyme via an ester bond to the serine's oxygen, forming a relatively stable acyl enzyme intermediate. The subsequent replacement of this oxygen with one from water, in a reaction involving another unstable tetrahedral intermediate, leads to release of the final product ( $\text{P}_1\text{—COOH}$ ). The tetrahedral intermediate transition states are partially stabilized by hydrogen bonding with the enzyme's backbone amino groups in what is called the *oxyanion hole*. The large family of serine proteases and



**FIGURE 3-28 Mechanism of serine protease–mediated hydrolysis of peptide bonds.** The catalytic triad of Ser-195, His-57, and Asp-102 in the active sites of serine proteases employs a multistep mechanism to hydrolyze peptide bonds in target proteins. (a) After a polypeptide substrate binds to the active site (see Figure 3-27), forming an ES complex, the hydroxyl oxygen of Ser-195 attacks the carbonyl carbon of the substrate's targeted peptide bond (yellow). Movements of electrons are indicated by arrows. (b) This attack results in the formation of a transition state called the *tetrahedral intermediate*, in which the negative charge on the substrate's oxygen is stabilized by hydrogen bonds formed with the enzyme's *oxyanion hole*. (c) Additional electron movements result in the breaking of the

peptide bond, release of one of the reaction products ( $\text{NH}_2\text{—P}_2$ ), and formation of the acyl enzyme ( $\text{ES}'$  complex). (d) An oxygen from a solvent water molecule then attacks the carbonyl carbon of the acyl enzyme. (e) This attack results in the formation of a second tetrahedral intermediate. (f) Additional electron movements result in the breaking of the Ser-195–substrate bond (formation of the EP complex) and release of the final reaction product ( $\text{P}_1\text{—COOH}$ ). The side chain of His-57, which is held in the proper orientation by hydrogen bonding to the side chain of Asp-102, facilitates catalysis by withdrawing and donating protons throughout the reaction (*inset*). If the pH is too low and the side chain of His-57 is protonated, it cannot participate in catalysis and the enzyme is inactive.



related enzymes, all of which have an active-site serine, illustrates how an efficient reaction mechanism is used over and over by distinct enzymes to catalyze similar reactions.

The serine protease mechanism points out several key features of enzymatic catalysis. First, enzyme catalytic sites have evolved to stabilize the binding of a transition state, thus lowering the activation energy and accelerating the overall reaction. Second, multiple side chains, together with the polypeptide backbone, carefully organized in three dimensions, work together to chemically transform substrate into product, often by multistep reactions.

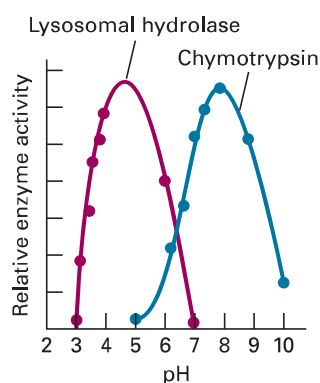
Third, acid-base catalysis mediated by one or more amino acid side chains is often used by enzymes, as when the imidazole group of His-57 in serine proteases acts as a base to remove the hydrogen from Ser-195's hydroxyl group. As a consequence, only a particular ionization state (protonated or nonprotonated) of one or more amino acid side chains in the catalytic site may be compatible with catalysis, and thus the enzyme's activity may be pH dependent. For example, the imidazole of His-57 in serine proteases, whose  $pK_a$  is  $\sim 6.8$ , can help the Ser-195 hydroxyl attack the substrate only if it is not protonated. Thus the activity of the protease is low at  $pH < 6.8$ , at which the imidazole is protonated, and the shape of the pH activity profile in the pH range 4–8 matches the titration of the His-57 side chain, which is governed by the Henderson-Hasselbalch equation, with an inflection near pH 6.8 (see chymotrypsin data in Figure 3-29 and see Chapter 2). The activity drops at higher pH values, generating a bell-shaped activity curve, because the proper folding of the protein is disrupted when the amino group at

the protein's amino terminus ( $pK_a \sim 9$ ) is deprotonated; the conformation near the active site changes as a consequence.

The pH sensitivity of an enzyme's activity can be due to changes in the ionization of catalytic groups, groups that participate directly in substrate binding, or groups that influence the conformation of the protein. Pancreatic serine proteases evolved to function in the neutral or slightly basic conditions in the intestines; hence their pH optima are  $\sim 8$ . Proteases and other hydrolytic enzymes that function in acidic conditions must employ a different catalytic mechanism. This is the case for enzymes within the stomach ( $pH \sim 1$ ), such as the protease pepsin, and for those within lysosomes ( $pH \sim 4.5$ ), which play a key role in degrading macromolecules within cells (see the lysosomal hydrolase data in Figure 3-29). Indeed, lysosomal hydrolases, which degrade a wide variety of biomolecules (proteins, lipids, etc.), are relatively inactive at the pH in the cytosol ( $\sim 7$ ), which helps to protect a cell from self-digestion if these enzymes escape the confines of the membrane-bounded lysosome.

One key feature of enzymatic catalysis not seen in serine proteases but found in many other enzymes is a *cofactor* or *prosthetic group*. This “helper” group is a nonpolypeptide small molecule or ion (e.g., iron, zinc, copper, manganese) that is bound in the active site and plays an essential role in the reaction mechanism. Small organic prosthetic groups in enzymes are also called *coenzymes*. Some of these groups are chemically modified during the reaction and thus need to be replaced or regenerated after each reaction; others are not. Examples include  $NAD^+$  (nicotinamide adenine dinucleotide), FAD (flavin adenine dinucleotide) (see Figure 2-33), and the heme groups that bind oxygen in hemoglobin or transfer electrons in some cytochromes (see Figure 12-20). Thus the chemical reactions catalyzed by enzymes are not restricted by the limited number of types of amino acids in polypeptide chains. Many of the vitamins [for example, the B vitamins thiamine ( $B_1$ ), riboflavin ( $B_2$ ), niacin ( $B_3$ ), and pyridoxine ( $B_6$ ), as well as vitamin C], which cannot be synthesized in mammalian cells, function as, or are used to generate, coenzymes. That is why supplements of vitamins must be added to the liquid medium in which mammalian cells are grown in the laboratory (see Chapter 4).


Small molecules that can bind to active sites and disrupt catalytic reactions are called *enzyme inhibitors*. Such inhibitors are useful tools for studying the roles of enzymes in cells and organisms. Inhibitors that bind directly to an enzyme's binding site and thus compete directly with the normal substrate are called competitive inhibitors. Noncompetitive inhibitors are those that interfere with enzyme activity in other ways—for example, by binding to some other site on the enzyme and changing its conformation. Enzyme inhibitors complement the use of genetic mutations and a technique called RNA interference (RNAi) for probing an enzyme's function in cells (see Chapter 6). In all three approaches, the cellular consequences of disrupting an enzyme's activity can be used to deduce the normal function of the enzyme. The same approaches can be used to study the functions of nonenzymatic macromolecules. Interpreting the results of inhibitor studies



**FIGURE 3-29 The pH dependence of enzyme activity.** In some cases, ionizable (pH-titratable) groups in enzyme active sites or elsewhere in enzymes must be either protonated or deprotonated to permit proper substrate binding or catalysis, or to permit the enzyme to adopt the correct conformation. Measurement of enzyme activity as a function of pH can be used to identify the  $pK_a$ 's of these groups. The pancreatic serine proteases, such as chymotrypsin, exhibit maximum activity at around pH 8 because of titration of the active-site His-57 (required for catalysis,  $pK_a \sim 6.8$ ) and of the amino terminus of the protein (required for proper conformation,  $pK_a \sim 9$ ). Many lysosomal hydrolases have evolved to exhibit a lower pH optimum ( $\sim 4.5$ ) to match the low internal pH in the lysosomes in which they function.

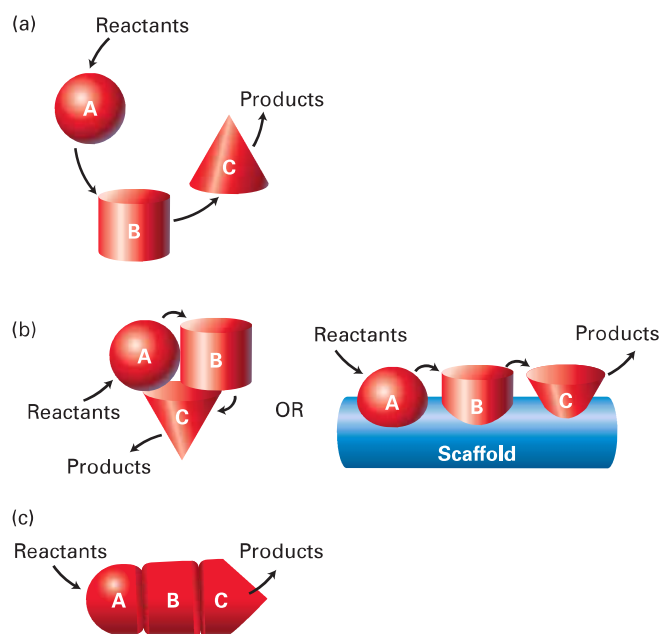
[Data from P. Lozano, T. De Diego, and J. L. Iborra, 1997, *Eur. J. Biochem.* **248**:80–85, and W. A. Judice et al., 2004, *Eur. J. Biochem.* **271**:1046–1053.]

can be complicated, however, if, as is often the case, the inhibitors block the activity of more than one protein.

 Small-molecule inhibition of protein activity is the basis for many drugs as well as for chemical warfare agents. Aspirin inhibits enzymes called cyclooxygenases, whose products can cause pain. Sarin and other nerve gases react with the active serine hydroxyl groups of both serine proteases and a related enzyme, acetylcholine esterase, which is a key enzyme in regulating nerve conduction (see Chapter 22). ■

### Enzymes in a Common Pathway Are Often Physically Associated with One Another

Enzymes taking part in a common metabolic process (e.g., the degradation of glucose to pyruvate during glycolysis; see Chapter 12) are generally located in the same cellular compartment, be it in the cytosol, at a membrane, or within a particular organelle. Within this compartment, products from one reaction can move by diffusion to the next enzyme in the metabolic pathway. Diffusion, however, entails random movement and can be a slow, relatively inefficient process for moving molecules between enzymes (Figure 3-30a).



**FIGURE 3-30 Assembly of enzymes into efficient multienzyme complexes.** In the hypothetical reaction pathways illustrated here, the initial reactants are converted into final products by the sequential action of three enzymes: A, B, and C. (a) When the enzymes are free in solution, or even constrained within the same cellular compartment, the intermediates in the reaction sequence must diffuse from one enzyme to the next, an inherently slow process. (b) Diffusion is greatly reduced or eliminated when the individual enzymes associate into multisubunit complexes, either by themselves or with the aid of a scaffold protein. (c) The closest integration of different catalytic activities occurs when the enzymes are fused at the genetic level, becoming domains in a single polypeptide chain.

To overcome this impediment, cells have evolved mechanisms for bringing enzymes in a common pathway into close proximity, a process called metabolic coupling.

In the simplest such mechanism, polypeptides with different catalytic activities cluster closely together as subunits of a multimeric enzyme or assemble on a common “scaffold” that holds them together (Figure 3-30b). This arrangement allows the product of one reaction to be channeled directly to the next enzyme in the pathway. In some cases, independent proteins have been fused together at the genetic level to create a single multidomain, multifunctional enzyme (Figure 3-30c). Metabolic coupling usually involves large multiprotein complexes, as described earlier in this chapter.

### KEY CONCEPTS OF SECTION 3.3

#### Protein Binding and Enzyme Catalysis

- A protein’s function depends on its ability to bind other molecules, known as ligands. For example, antibodies bind to a group of ligands known as antigens, and enzymes bind to reactants called substrates that will be converted by chemical reactions into products.
- The specificity of a protein for a particular ligand refers to the preferential binding of one or a few closely related ligands. The affinity of a protein for a particular ligand refers to the strength of binding, usually expressed as the dissociation constant  $K_d$ .
- Proteins are able to bind to ligands because of molecular complementarity between the ligand-binding sites and the corresponding ligands.
- Enzymes are catalytic proteins that accelerate the rates of cellular reactions by lowering the activation energy and stabilizing transition-state intermediates (see Figure 3-22).
- An enzyme’s active site, which is usually only a small part of the protein, comprises two functional parts: a substrate-binding site and a catalytic site. The substrate-binding site is responsible for the exquisite specificity of enzymes owing to its molecular complementarity with the substrate.
- The initial binding of a substrate (S) to an enzyme (E) results in the formation of an enzyme-substrate complex (ES), which then undergoes one or more reactions catalyzed by the catalytic groups in the catalytic site until the final product (P) is formed.
- From plots of reaction rate versus substrate concentration, two characteristic parameters of an enzyme can be determined: the Michaelis constant,  $K_m$ , a rough measure of the enzyme’s affinity for converting substrate into product, and the maximal velocity,  $V_{max}$ , a measure of its catalytic power (see Figure 3-24).
- The rates of enzyme-catalyzed reactions vary enormously, with turnover numbers (numbers of substrate molecules converted to product at a single active site at substrate saturation) ranging from fewer than 1 to  $6 \times 10^5$  molecules per second.

- Many enzymes catalyze the conversion of substrates to products by dividing the process into multiple discrete chemical reactions that involve multiple distinct enzyme-substrate complexes (ES', ES'' etc.).
- Serine proteases hydrolyze peptide bonds in substrate proteins using as catalytic groups the side chains of Ser-195, His-57, and Asp-102. Amino acids lining the side-chain-specificity binding pocket in the binding site of serine proteases determine the residue in a substrate protein whose peptide bond will be hydrolyzed and account for differences in protease specificity (for example, trypsin vs. chymotrypsin and elastase).
- Enzymes often use acid-base catalysis mediated by one or more amino acid side chains, such as the imidazole group of His-57 in serine proteases, to catalyze reactions. The pH dependence of protonation of catalytic groups ( $pK_a$ ) is often reflected in the pH-rate profile of the enzyme's activity.
- Nonpolypeptide small molecules or ions, called cofactors or prosthetic groups, bind to the active sites of some enzymes and play an essential role in enzymatic catalysis. Small organic prosthetic groups in enzymes are also called coenzymes; many vitamins, which cannot be synthesized in higher animal cells, function as or are used to generate coenzymes.
- Enzymes in a common metabolic pathway are often located within the same cellular compartments and may be further associated as domains of a monomeric protein, subunits of a multimeric protein, or components of a protein complex assembled on a common scaffold (see Figure 3-30).

### 3.4 Regulating Protein Function

Most processes in cells do not take place independently of one another or at a constant rate. The activities of all proteins and other biomolecules are regulated to integrate their functions for optimal performance for survival. For example, the catalytic activity of enzymes is regulated so that the amount of reaction product is just sufficient to meet the needs of the cell. As a result, the steady-state concentrations of substrates and products may vary depending on cellular conditions. Regulation of nonenzymatic proteins—the opening or closing of membrane channels or the assembly of a macromolecular complex, for example—is also essential.

In general, there are three ways to regulate protein activity. First, cells can increase or decrease the steady-state level of the protein by altering its rate of synthesis, its rate of degradation, or both. Second, cells can change the intrinsic activity, as distinct from the amount, of the protein. For example, through noncovalent and covalent interactions, cells can change the affinity of substrate binding, or the fraction of time the protein is in an active versus an inactive conformation. Third, there can be a change in the location or the concentration within the cell of the protein itself, of the target of the protein's activity (e.g., an enzyme's substrate), or of some other molecule required for the protein's

activity (e.g., an enzyme's cofactor). All three types of regulation play essential roles in the lives and functions of cells. In this section, we first discuss mechanisms for regulating the amount of a protein, then turn to noncovalent and covalent interactions that regulate protein activity.

### Regulated Synthesis and Degradation of Proteins Is a Fundamental Property of Cells

The rate of synthesis of a protein is determined by the rate at which the DNA encoding the protein is converted to mRNA (transcription), the steady-state amount of the active mRNA in the cell, and the rate at which the mRNA is converted into newly synthesized protein (translation). These important processes are described in detail in Chapter 5.

The life spans of intracellular proteins vary from as short as a few minutes for mitotic cyclins, which help regulate passage through the mitotic stage of cell division (see Chapter 19), to as long as the age of an organism for proteins in the lens of the eye. Protein life span is controlled primarily by regulated protein degradation.

Protein degradation plays two especially important roles in the cell. First, it removes proteins that are potentially toxic, improperly folded or assembled, or damaged—including the products of mutated genes and proteins damaged by chemically active cell metabolites or stress (e.g., heat shock). Despite the existence of chaperone-mediated protein folding, some newly made proteins are rapidly degraded because they are misfolded. This degradation might be necessary due to failure of timely engagement of the necessary chaperones to guide the folding of the proteins or due to their defective assembly into complexes. Most other proteins are degraded more slowly, undergoing about 1–2 percent degradation per hour in mammalian cells. Second, the controlled destruction of otherwise normal proteins, along with controlled rates of synthesis, provides a powerful mechanism for maintaining the appropriate levels of the proteins and their activities and for permitting rapid changes in these levels to help the cells respond to changing conditions.

Eukaryotic cells have several pathways for degrading proteins. One major pathway is degradation by enzymes within lysosomes, membrane-limited organelles whose acidic interior (pH ~4.5) is filled with a host of hydrolytic enzymes. Lysosomal degradation is directed primarily toward aged or defective organelles of the cell—a process called autophagy (see Chapter 14)—and toward extracellular proteins taken up by the cell. Lysosomes will be discussed at length in later chapters. Here we focus on another important degradation pathway: cytoplasmic protein degradation by proteasomes, which can account for up to 90 percent of the protein degradation in mammalian cells.

### The Proteasome Is a Molecular Machine Used to Degrade Proteins

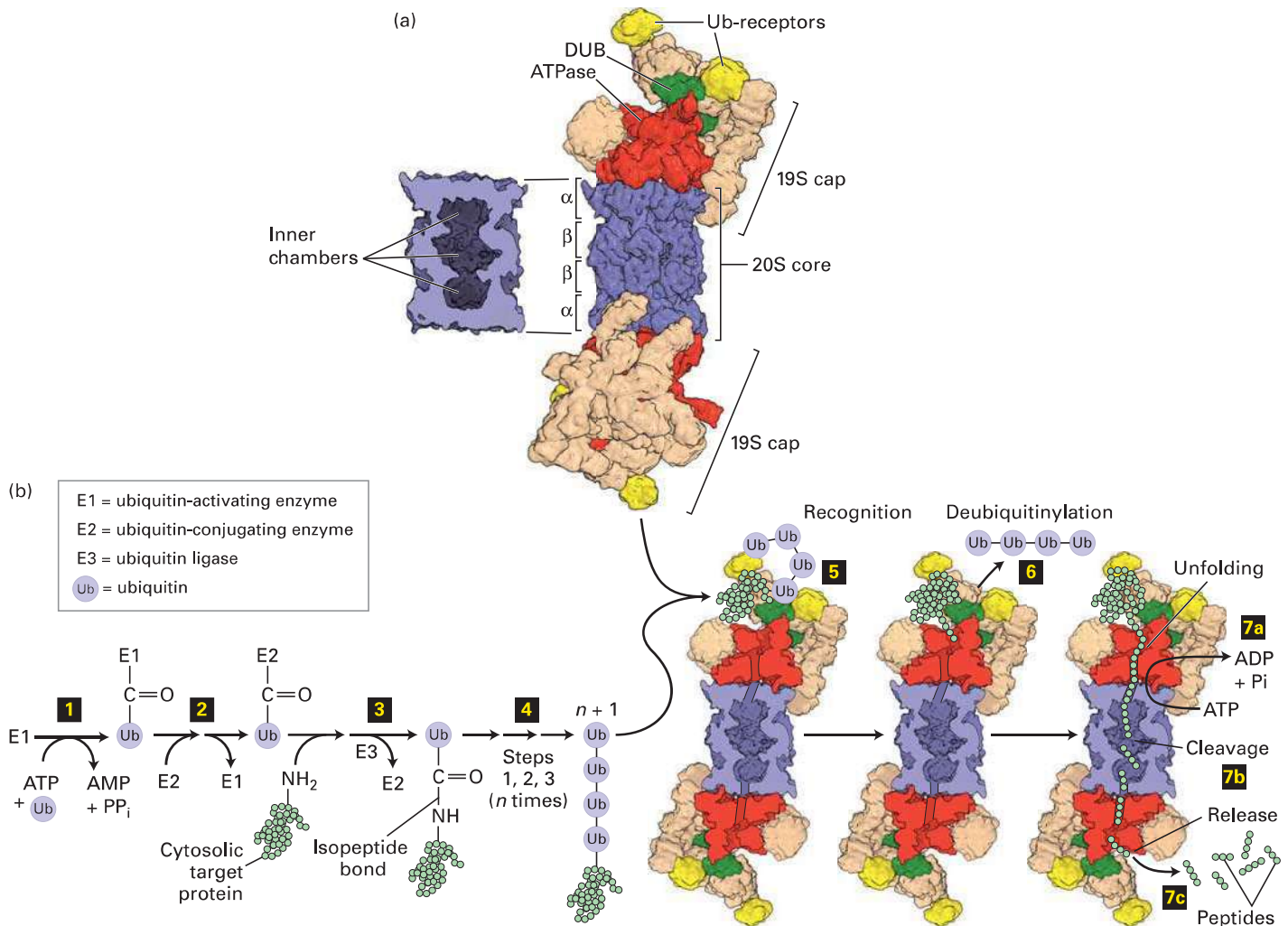
**Proteasomes** are very large protein-degrading molecular machines that influence many different cellular functions,



including the cell cycle (see Chapter 19), transcription, DNA repair (see Chapter 5), programmed cell death, or **apoptosis** (see Chapter 21), recognition of and response to infection by foreign organisms (see Chapter 23), and removal of misfolded proteins. There are approximately 30,000 proteasomes in a typical mammalian cell.

Proteasomes consist of roughly 60 protein subunits and have a mass of about  $2.4 \times 10^6$  Da. Proteasomes have a

cylindrical, barrel-like catalytic core (Figure 3-31a), called the *20S proteasome* (where *S* is a Svedberg unit based on the sedimentation properties of the particle and is proportional to its size), which is approximately 14.8 nm tall and 11.3 nm in diameter. Bound to the ends of this core are either one or two 19S cap complexes that regulate the activity of the 20S catalytic core. When the core and one or two caps are combined, they are referred to collectively as the *26S complex*,



**FIGURE 3-31 Ubiquitin- and proteasome-mediated proteolysis.**

(a) *Right*: The 26S proteasome has a cylindrical structure with a 19S cap at one or both ends of the 20S core particle. The nineteen different subunits of the 19S cap (shown in multiple colors) include six AAA-ATPase subunits (Rpt1–6, red), which assemble into a heterohexameric ring; two ubiquitin (Ub) receptors (Rpn10 and Rpn13, yellow); and a deubiquitinase enzyme (DUB, Rpn11, green), which forms a heterodimer with its evolutionarily related counterpart Rpn8. Moreover, the 19S cap contains scaffolding and other proteins (tan). The two 19S caps shown are facing in opposite directions relative to the plane of the page. The 20S core consists of four stacked heptameric rings (~110 Å diameter × 160 Å long), each containing either α (outer rings) or β (inner rings) subunits (blue). *Left*: Cutaway view of the 20S core, showing the inner chambers. Proteolysis occurs within the central inner chamber of the core formed by the β rings.

(step **1**) and then transfers this Ub molecule to a cysteine residue in E2 (step **2**). Ubiquitin ligase (E3) transfers the bound Ub molecule on E2 to the side-chain –NH<sub>2</sub> of a lysine residue in a target protein, forming an isopeptide bond (step **3**). Additional Ub molecules are added to the Ub-modified target protein via isopeptide bonds to the previously added Ub by repeating steps **1** – **3**, forming a polyubiquitin chain (step **4**). The polyubiquitylated target is recognized by Ub receptors in the proteasome's 19S cap (step **5**), and the Ub groups are removed by the deubiquitinase enzyme (step **6**). In step **7** ATP hydrolysis enables the six protein (hexameric) ATPase subunits (red) to unfold the substrate and transfer the unfolded protein via a pore in the hexamer into the proteolysis chamber in the 20S core (step **7a**), in some cases coordinately with step **6**, and the protein is cleaved into short peptide digestion fragments (step **7b**) that are then released (step **7c**). [Part (a) courtesy of Antje Aufderheide and Friedrich Foerster, data from P. Unverdorben et al., 2014, *P. Natl. Acad. Sci. USA* **111**(15):5544–5549, PDB ID 4cr2.]

even though the two-cap-containing complex is larger (30S). A 19S cap has 19 protein subunits, six of which can hydrolyze ATP (i.e., they are AAA-type ATPases) to provide the energy needed to unfold protein substrates and selectively transfer them into the inner chamber of the proteasome's catalytic core. Genetic studies in yeast have shown that cells cannot survive without functional proteasomes, thus demonstrating their importance. Furthermore, proper proteasomal activity is so important that cells will expend as much as 30 percent of the energy needed to synthesize a protein to degrade it in a proteasome.

The 20S proteasomal catalytic core comprises two inner rings of seven  $\beta$  subunits each, with three proteolytic active sites per ring facing toward the  $\sim 1.7$ -nm-diameter inner chamber formed by those rings, and two outer rings of seven  $\alpha$  subunits each, which limit substrate access (see Figure 3-31a) via an entry channel that can be opened by the 19S cap. Proteasomes can degrade most proteins thoroughly because the three active sites in each  $\beta$  subunit ring can cleave peptide bonds at hydrophobic residues, acidic residues, or basic residues. Polypeptide substrates must enter the chamber via a regulated  $\sim 1.3$ -nm-diameter aperture at the center of the outer  $\alpha$  subunit rings. In the 26S proteasome, the opening of the aperture, which is narrow and often allows the entry of only unfolded proteins, is controlled by ATPases in the 19S cap. These ATPases are responsible for unfolding protein substrates and translocating those unfolded polypeptides into the inner chamber of the catalytic core (Figure 3-31b, *bottom right*). The short peptide products of proteasomal digestion (2–24 residues long) exit the chamber and are further degraded rapidly by cytosolic peptidases, eventually being converted to individual (“free”) amino acids. One researcher has quipped that a proteasome is a “cellular chamber of doom” in which proteins suffer a “death by a thousand cuts.”



Inhibitors of proteasome function have proved to be exceptionally useful in the laboratory and the clinic. Small-molecule proteasome inhibitors, such as MG132, are used to block proteasomal degradation in the lab and to help evaluate the role of the proteasome and, as we shall see below, polyubiquitinylation in a wide variety of processes. Other small-molecule proteasome inhibitors have been used therapeutically. Because of the global importance of proteasome-mediated protein breakdown in cells, continuous, complete inhibition of proteasomes kills cells. However, partial proteasome inhibition for short intervals is widely used as an approach to cancer chemotherapy, especially to treat multiple myeloma, a cancer involving the abnormal proliferation of antibody (immunoglobulin)-producing cells. The myeloma cells produce abnormally high levels of potentially toxic, aberrant immunoglobulin polypeptide chains, which are degraded by proteasomes. Proteasome inhibition in these cancer cells leads to the buildup of toxic, misfolded immunoglobulin polypeptides within the cells, and thus to cell death. In addition, to survive and grow, myeloma cells require the robust activity of a regulatory protein called *NF- $\kappa$ B* (see Chapter 16) as well as other “pro-survival” and

“pro-proliferation” proteins. In turn, *NF- $\kappa$ B* can function fully and promote survival and proliferation only when its inhibitor, *I- $\kappa$ B*, is disengaged and degraded by proteasomes (see Chapter 16). Partial inhibition of proteasomal activity by a small-molecule inhibitor drug results in increased levels of *I- $\kappa$ B* and, consequently, reduced *NF- $\kappa$ B* activity (that is, loss of its protective activity). The cancer cells subsequently undergo less proliferation and die by apoptosis. Thus, multiple myeloma cells are more sensitive to proteasome inhibitors than normal cells. Consequently, *controlled* administration of proteasome inhibitors, at levels that kill the cancer cells but not normal cells, has proved to be an effective therapy for multiple myeloma. ■

## Ubiquitin Marks Cytosolic Proteins for Degradation in Proteasomes

If proteasomes are to rapidly degrade only those proteins that are either defective or scheduled to be removed, they must be able to distinguish between those proteins that need to be degraded and those that don't. Cells mark proteins that should be degraded by covalently attaching to them a linear chain of multiple copies of a 76-residue polypeptide called *ubiquitin* (Ub) that is highly conserved from yeast to humans. This “polyubiquitin tail” serves as a cellular “kiss of death,” marking the protein for destruction in the proteasome. The ubiquitinylation process (Figure 3-31b, steps 1–3) involves three distinct steps:

1. Activation of *ubiquitin-activating enzyme* (E1) by the addition of a ubiquitin molecule, a reaction that requires ATP.
2. Transfer of this ubiquitin molecule to a cysteine residue in a *ubiquitin-conjugating enzyme* (E2).
3. Formation of a covalent bond between the carboxyl group of the C-terminal glycine 76 of the ubiquitin bound to E2 and the amino group of the side chain of a lysine residue in the target protein, a reaction catalyzed by a *ubiquitin-protein ligase* (E3). This type of bond is called an *isopeptide bond* because it covalently links a side-chain amino group, rather than the  $\alpha$  amino group, to the carboxyl group. Subsequent ligase reactions covalently attach the C-terminal glycine of an additional ubiquitin molecule via an isopeptide bond to the side chain of lysine 48 of the previously added ubiquitin to generate a polyubiquitin chain covalently attached to the target protein. (We will discuss ubiquitin linkages via other lysine side chains shortly.)

Generally, following attachment of four or more ubiquitins in a polyubiquitin chain, the 19S regulatory cap of the 26S proteasome (sometimes with the help of accessory proteins) recognizes the polyubiquitin-labeled protein using its Ub receptors (see Figure 3-31a), uses ATPases to unfold it, and transports it into the proteasome core for degradation. As a polyubiquitinated substrate is unfolded and passed into the core of the proteasome, enzymes called deubiquitinases (Dubs) hydrolyze the bonds between the individual ubiquitins and between the targeted protein and ubiquitin,

recycling the ubiquitins for additional rounds of protein modification (see Figure 3-31b). Analysis of the human genome sequence indicates the presence of about 90 distinct Dubs, about 80 percent of which use cysteine in a catalytic triad similar to that in the serine proteases described earlier (the sulfhydryl in the cysteine side chain is used in place of the hydroxyl in the side chain of the serine). In some Dubs, zinc is a key participant in the catalytic reactions.

**Specificity of Degradation** Targeting of specific proteins for proteasomal degradation is primarily achieved through the substrate specificity of E3 ligases (see Figure 3-31b, step 3). As a testament to their importance, there are an estimated 600 or more ubiquitin ligase genes in the human genome. The many E3 ligases in mammalian cells ensure that the wide variety of proteins to be polyubiquitinated can be modified when necessary. Some E3 ligases are associated with chaperones that recognize unfolded or misfolded proteins; for example, the E3 ligase CHIP is a co-chaperone for Hsp70. These and other proteins (co-chaperones, escort factors, adapters) can mediate E3 ligase-catalyzed polyubiquitination of dysfunctional proteins that cannot be readily refolded properly and, consequently, mediate their delivery to proteasomes for degradation. In such cases, the chaperone-ubiquitination-proteasome system works in concert for protein quality control.

In addition to quality control, the ubiquitin-proteasome system can be used to regulate the activity of important cellular proteins. An example is the regulated degradation of proteins called *cyclins*, which control the cell cycle (see Chapter 19). Cyclins contain the internal sequence Arg-X-X-Leu-Gly-X-Ile-Gly-Asp/Asn (where X can be any amino acid), which is recognized by specific ubiquitinating enzyme complexes. At a specific time in the cell cycle, each cyclin is phosphorylated by a cyclin kinase. This phosphorylation is thought to cause a conformational change that exposes the recognition sequence to the ubiquitinating enzymes, leading to polyubiquitination and proteasomal degradation.

**Other Functions of Ubiquitin and Ubiquitin-Related Molecules** There are several close relatives of ubiquitin that employ similar E1-, E2-, and E3-dependent mechanisms of activation and transfer to acceptor substrates. These ubiquitin-like modifiers control processes as diverse as nuclear import, regulated by the ubiquitin-like modifier Sumo, and autophagy, regulated by the ubiquitin-like modifier Atg8/LC3 (see Chapter 14). Furthermore, the attachment of ubiquitin to a target protein can be used for purposes other than to mark the protein for degradation, as we will see later in this section, and some of these functions involve polyubiquitin linkages other than those via Lys-48.

Like ubiquitination, deubiquitination is involved in processes other than proteasome-mediated protein degradation. Large-scale, mass-spectrometry-based “proteomic” methods described later in this chapter, together with sophisticated computational approaches, have suggested that Dubs, which are often bound in multiprotein complexes, are involved in an extraordinarily wide range of cell processes.

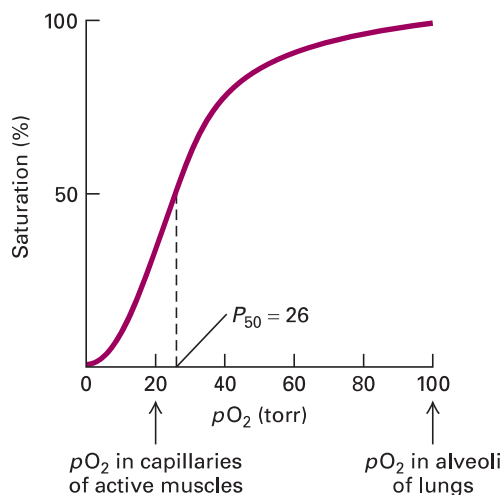
These processes vary from cell division and cell cycle control (see Chapter 19) to membrane trafficking (see Chapter 14) to cell signaling pathways (see Chapters 15 and 16).

## Noncovalent Binding Permits Allosteric, or Cooperative, Regulation of Proteins

In addition to regulating the amount of a protein, cells can also regulate the intrinsic activity of a protein. One of the most important mechanisms for regulating protein function is through allosteric interactions. Broadly speaking, **allostery** (from the Greek, “other shape”) refers to any change in a protein’s tertiary or quaternary structure, or in both, induced by the noncovalent binding of a ligand. When a ligand binds to one site (A) in a protein and induces a conformational change that alters the activity of a different site (B), the ligand is called an *allosteric effector* of the protein, while site A is called an *allosteric binding site*, and the protein is called an *allosteric protein*. By definition, allosteric proteins have multiple binding sites, at least one for the allosteric effector and at least one for other molecules with which the protein interacts. The allosteric change in activity can be positive or negative; that is, it can be an increase or a decrease in protein activity. Negative allostery often involves the end product of a multi-step biochemical pathway binding to, and reducing the activity of, an enzyme that catalyzes an early, rate-controlling step in that pathway. In this way, excessive buildup of the product is prevented. This kind of regulation of a metabolic pathway is also called *end-product inhibition* or *feedback inhibition*. Allosteric regulation is particularly prevalent in multimeric enzymes and other proteins in which conformational changes in one subunit are transmitted to an adjacent subunit.

*Cooperativity*, a term that is often used synonymously with *allostery*, usually refers to the influence (positive or negative) that the binding of a ligand at one site has on the binding of another molecule of the *same* type of ligand at a different site. Hemoglobin presents a classic example of positive cooperative binding in that the binding of a single ligand, molecular oxygen ( $O_2$ ), increases the affinity of hemoglobin for the next oxygen molecule. Each of the four subunits in hemoglobin contains one heme molecule. The heme groups are the oxygen-binding components of hemoglobin (see Figure 3-14a). The binding of oxygen to the heme molecule in one of the four hemoglobin subunits induces a local conformational change whose effect spreads to the other subunits, lowering the  $K_d$  (increasing the affinity) for the binding of additional oxygen molecules to the remaining hemes and yielding a sigmoidal oxygen-binding curve (Figure 3-32). Because of the sigmoidal shape of the oxygen-saturation curve, it takes only a fourfold increase in oxygen concentration for the saturation of the oxygen-binding sites in hemoglobin to go from 10 to 90 percent. Conversely, if there were no cooperativity and the shape of the curve was typical of that for Michaelis-Menten (see Figure 3-24), or noncooperative, binding, it would take an eighty-one-fold increase in oxygen concentration to accomplish the same increase in loading of its binding sites in hemoglobin. This cooperativity permits hemoglobin to take up oxygen very efficiently in the lungs, where the oxygen





**EXPERIMENTAL FIGURE 3-32 Hemoglobin binds oxygen cooperatively.** Each tetrameric hemoglobin molecule has four oxygen-binding sites; at saturation, all the sites are loaded with oxygen. The oxygen concentration in tissues is commonly measured as the partial pressure of oxygen ( $pO_2$ ) in torr units (a standard measure of pressure equivalent to 1 mm of mercury under standard conditions).  $P_{50}$  is the  $pO_2$  at which half the oxygen-binding sites are occupied; it is somewhat analogous to the  $K_m$  for an enzymatic reaction. The large change in the amount of oxygen bound over a small range of  $pO_2$  values permits efficient unloading of oxygen in peripheral tissues such as muscle. The sigmoidal shape of a plot of saturation versus ligand concentration is indicative of cooperative binding, in which the binding of one oxygen molecule allosterically influences the binding of subsequent oxygens. In the absence of cooperative binding, a binding curve is a hyperbola, similar to the curves in Figure 3-24. See J. M. Berg et al., 2015, *Biochemistry*, 8th ed., Macmillan.

concentration is high, and unload it in tissues, where the concentration is low. Thus cooperativity amplifies the sensitivity of a system to changes in the concentration of its ligands, providing in many cases a selective evolutionary advantage.

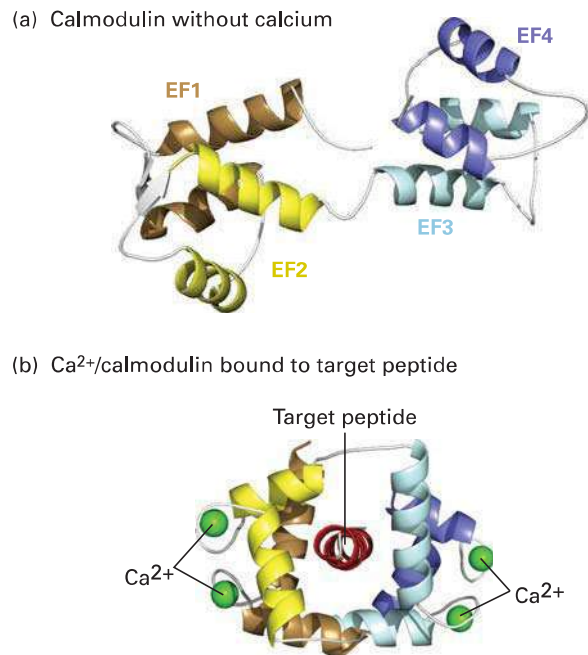
## Noncovalent Binding of Calcium and GTP Are Widely Used as Allosteric Switches to Control Protein Activity

Unlike oxygen, which causes graded allosteric changes in the activity of hemoglobin, some other allosteric effectors act as switches, turning the activity of many different proteins on or off by binding to them noncovalently. Two important allosteric switches that we will encounter many times throughout this book, especially in the context of cell signaling pathways (see Chapters 15 and 16), are  $Ca^{2+}$  and GTP.

**$Ca^{2+}$ /Calmodulin-Mediated Switching** The concentration of  $Ca^{2+}$  that is free in the cytosol (not bound to molecules other than water) is kept very low ( $\sim 10^{-7}$  M) by specialized membrane transport proteins that continually pump excess  $Ca^{2+}$  out of the cytosol (see Chapters 11 and 15). However, as we will learn in Chapters 11 and 15, the cytosolic  $Ca^{2+}$  concentration can increase tenfold to a hundredfold when  $Ca^{2+}$ -permeable channels in the cell-surface membranes open

and allow extracellular  $Ca^{2+}$  to flow into the cell. This rise in cytosolic  $Ca^{2+}$  is sensed by specialized  $Ca^{2+}$ -binding proteins, which alter cellular behavior by turning the activities of other proteins on or off. The importance of extracellular  $Ca^{2+}$  for cell activity was first documented by S. Ringer in 1883, when he discovered that isolated rat hearts suspended in an NaCl solution made with “hard” ( $Ca^{2+}$ -rich) London tap water contracted beautifully, whereas they beat poorly and stopped quickly if distilled,  $Ca^{2+}$ -depleted, water was used.

Many  $Ca^{2+}$ -binding proteins bind  $Ca^{2+}$  using the EF hand/helix-loop-helix structural motif discussed earlier (see Figure 3-10b). A well-studied EF hand protein, **calmodulin**, is found in all eukaryotic cells, where it may exist as an individual monomeric protein or as a subunit of a multimeric protein. This dumbbell-shaped molecule contains four  $Ca^{2+}$ -binding EF hands with  $K_d$ s of about  $10^{-6}$  M. The binding of  $Ca^{2+}$  to calmodulin causes a conformational change that permits  $Ca^{2+}$ /calmodulin to bind to conserved sequences in various target proteins (Figure 3-33), thereby switching their activities on or off. Calmodulin and similar EF hand proteins thus function as *switch proteins*, acting in concert with changes in  $Ca^{2+}$  levels to modulate the activity of other proteins.




**FIGURE 3-33 Conformational changes induced by  $Ca^{2+}$  binding to calmodulin.** Calmodulin is a widely distributed cytosolic protein that contains four  $Ca^{2+}$ -binding sites, one in each of its EF hand (helix-loop-helix) motifs (EF1-ER4, see also Figure 3-10). At cytosolic  $Ca^{2+}$  concentrations above about  $5 \times 10^{-7}$  M, binding of  $Ca^{2+}$  to calmodulin changes the protein’s conformation from the dumbbell-shaped,  $Ca^{2+}$ -free form (a) to one in which hydrophobic side chains become more exposed to solvent. The resulting  $Ca^{2+}$ /calmodulin complex can wrap around exposed helices (target peptides) with specialized sequences in various target proteins (b), thereby altering their activities. [Part (a) data from H. Kuboniwa et al., 1995, *Nat. Struct. Biol.* **2**:768–776, PDB ID 1cfd. Part (b) data from W. E. Meador, A. R. Means, and F. A. Quiocho, 1992, *Science* **257**:1251–1255, PDB ID 1cdl.]

**Switching Mediated by Guanine Nucleotide-Binding Proteins** Another group of intracellular switch proteins constitutes the **GTPase superfamily**. As the name suggests, these proteins are enzymes—GTPases—that can hydrolyze GTP (guanosine triphosphate) to GDP (guanosine diphosphate). They include the monomeric Ras protein (whose structure is shown in Figure 3-9, with bound GDP shown in blue) and the  $G_\alpha$  subunit of the trimeric G proteins, both discussed at length in Chapters 15 and 16. Both Ras and  $G_\alpha$  can bind to the plasma membrane, function in cell signaling, and play key roles in cell proliferation and differentiation. Other members of the GTPase superfamily function in protein synthesis, the transport of proteins between the nucleus and the cytoplasm, the formation of coated vesicles and their fusion with target membranes, and rearrangements of the actin cytoskeleton. Some GTPase proteins have a covalently attached lipid chain (see Figure 7-19) that mediates their binding to membranes. We examine the roles of various GTPase switch proteins in regulating intracellular signaling and other processes in several later chapters.

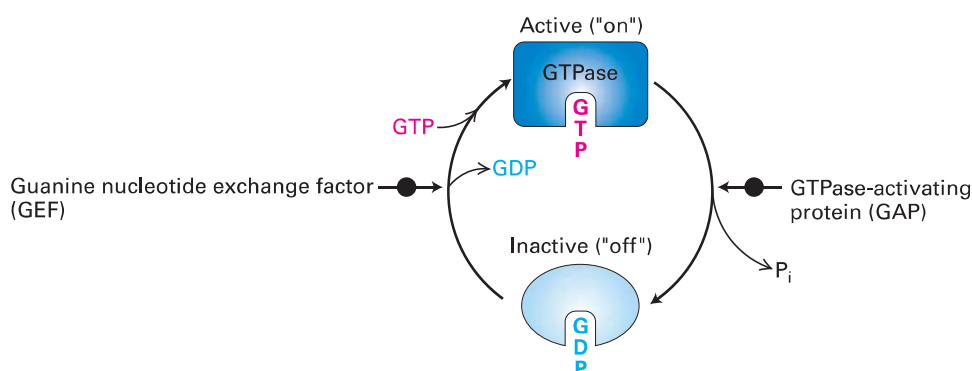
All the GTPase switch proteins exist in two forms, or conformations (Figure 3-34): (1) an active (“on”) form with bound GTP, which can influence the activity of specific target proteins to which they bind, and (2) an inactive (“off”) form with bound GDP. The switch is turned on—that is, the conformation of the protein changes from inactive to active—when a GTP molecule replaces a bound GDP in the inactive conformation. The switch is turned off when the relatively slow GTPase activity of the protein hydrolyzes bound GTP, converting it to GDP and leading the conformation to change to the inactive form. The amount of time any given GTPase switch remains in the active, GTP-bound form depends on the rate of its GTPase activity. Thus the GTPase activity acts as a timer to control this switch. Cells contain a variety of proteins that can modulate the baseline (or intrinsic) rate of GTPase activity for any given GTPase switch and so can control how long the switch remains on.

For example, GTPase-activating proteins, or GAPs, increase the rate of GTPase activity, thus reducing the time the GTPase is in the active form. Cells also have specific proteins whose function is to regulate the conversion of inactive GTPases to active ones—that is, to turn the switch on—by mediating the replacement of bound GDP with GTP (*GDP/GTP exchange*). These proteins are called guanine nucleotide exchange factors, or GEFs. GTPases with lipid anchors are also regulated by proteins called guanine nucleotide dissociation inhibitors (GDIs) that bind to the lipid chain and thus influence interactions with cellular membranes.

 The GAPs, GEFs, and GDIs are themselves subject to regulation and, together with their GTPases, participate in complex regulatory networks that control a vast array of cellular activities. It is, therefore, not surprising that disruptions of these finely tuned regulatory networks by mutations or pathogens are associated with a wide variety of diseases. Examples of genetic diseases affecting these networks include Noonan syndrome (a developmental disorder), retinitis pigmentosa (a degenerative eye disease), and X-linked mental retardation. Examples of disruptions of these networks by pathogens include bacterially induced food poisoning, dysenteries (inflammation of the intestines with diarrhea), Legionnaires’ disease (a severe type of pneumonia that involves lung inflammation), and even the plague [also called the Black Death, which between 1347 and 1351 decimated the populations of China (~50 percent death rate) and Europe (~33 percent death rate)]. ■

### Phosphorylation and Dephosphorylation Covalently Regulate Protein Activity

In addition to exploiting the noncovalent regulators described above, cells can use covalent modifications to regulate the intrinsic activity of a protein. One of the most common covalent mechanisms for regulating protein activity

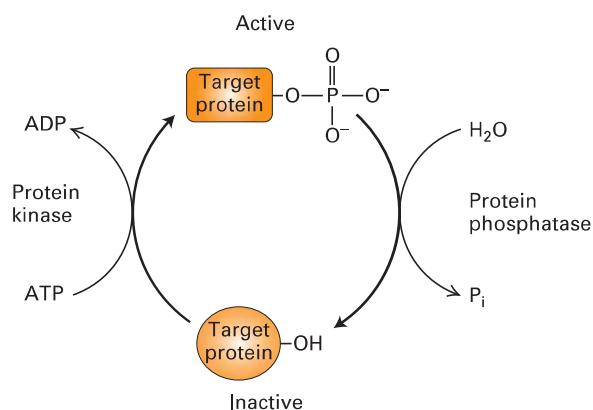


**FIGURE 3-34 The GTPase switch.** GTPases are enzymes that bind to GTP and hydrolyze it to GDP. When bound to GTP, the GTPase protein adopts its active, or “on,” conformation and can interact with target proteins to regulate their activities. When the bound GTP is hydrolyzed to GDP by the intrinsic GTPase activity of the protein, the GTPase with GDP bound assumes an inactive, or “off,” conformation.

The GTPase switch can be turned back on when another protein, called a GEF (guanine nucleotide exchange factor), mediates the replacement (exchange) of the bound GDP with a GTP molecule from the surrounding fluid. GTPase-activating proteins, or GAPs, can influence the rates of GTP hydrolysis. The binding of the active form of the GTPase to its targets is a form of noncovalent regulation.

is **phosphorylation**, the reversible addition of phosphate groups to hydroxyl groups on the side chains of serine, threonine, or tyrosine residues. Phosphorylated proteins are called *phosphoproteins*. Phosphorylation is catalyzed by enzymes called **protein kinases**, while the removal of phosphates, known as *dephosphorylation*, is catalyzed by **phosphatases**. The counteracting activities of kinases and phosphatases provide cells with a “switch” that can turn on or turn off the function of various proteins that are the substrates (or targets) of these enzymes (Figure 3-35). Sometimes phosphorylation sites are masked transiently by reversible covalent modification with the sugar N-acetylglucosamine (called O-GlcNAcylation), which is an additional means of covalent regulation. Phosphorylation changes a protein’s charge and to some extent its surface shape; it can also result in conformational changes. As a consequence, phosphorylation (or dephosphorylation) can influence the location of a protein within cells (e.g., its attachment to the inner surface of the plasma membrane), its intrinsic (e.g., enzymatic) activity, its ability to bind to other molecules, including metabolites, DNA, or other proteins, its ability to undergo further covalent modification, or its stability (rate of degradation). In addition, several conserved protein domains, such as the SH2 domain (see Figure 16-11), bind specifically to phosphorylated peptides. Thus phosphorylation can mediate the formation of protein complexes that can generate or extinguish a wide variety of cellular activities, discussed in many subsequent chapters.

Nearly 3 percent of all yeast proteins are protein kinases or phosphatases, indicating the importance of phosphorylation and dephosphorylation reactions even in these simple cells. Analysis of the human genome indicates there are approximately 500 human protein kinases (the human “kinome”). All classes of proteins—including structural proteins, scaffolds, enzymes, membrane channels, and signaling molecules—have members regulated by kinase/phosphatase modifications. Different protein kinases and phosphatases



**FIGURE 3-35 Regulation of protein activity by phosphorylation and dephosphorylation.** The cyclic phosphorylation and dephosphorylation of a protein is a common cellular mechanism for regulating protein activity. In this example, the target protein is active (*top*) when phosphorylated and inactive (*bottom*) when dephosphorylated; some proteins have the opposite response to phosphorylation.

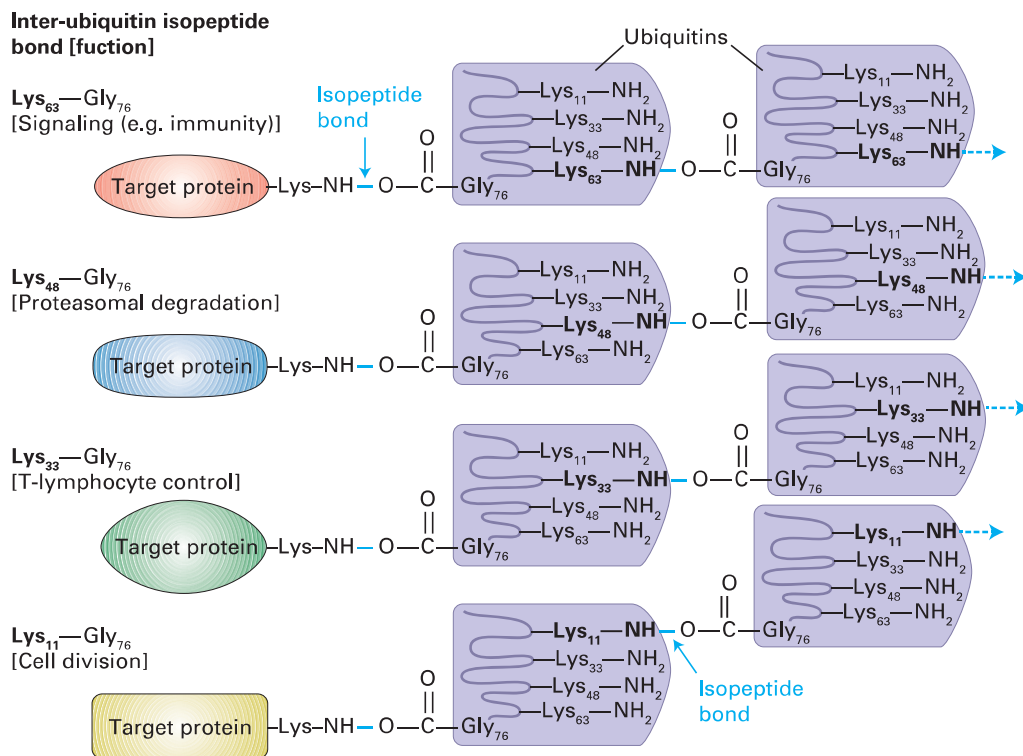
are specific for different target proteins, often recognizing different linear sequences in which the residue to be phosphorylated is embedded, and so can regulate distinct cellular pathways, as discussed in later chapters. Some kinases have many targets, so that a single kinase can serve to integrate the activities of many targets simultaneously. Frequently, the target of a kinase or phosphatase is yet another kinase or phosphatase, creating a cascade effect. There are many examples of such kinase cascades, which permit amplification of a signal and many levels of fine-tuning (see Chapters 15 and 16).

## Ubiquitinylation and Deubiquitinylation Covalently Regulate Protein Activity

Both ubiquitin and ubiquitin-like proteins (of which there are more than a dozen in humans) can be covalently linked to a target protein in a regulated fashion, in a manner analogous to phosphorylation. Deubiquitinases can reverse ubiquitinylation in a manner analogous to the action of phosphatases. These ubiquitin modifications are structurally far more complex than phosphorylation, however, and so can mediate many distinct interactions between the ubiquitinylated protein and other cellular proteins. Ubiquitinylation can involve attachment of a single ubiquitin to a protein (**monoubiquitinylation**), addition of multiple, single ubiquitin molecules to different sites on one target protein (**multiubiquitinylation**), or addition of a polymeric chain of ubiquitins to a protein (**polyubiquitinylation**). An additional source of variation is that different amino groups in the ubiquitin molecule can be used to form an isopeptide bond with the C-terminal Gly-76 in another ubiquitin to form a polyubiquitin chain. All seven lysine residues in ubiquitin (Lys-6, Lys-11, Lys-27, Lys-29, Lys-33, Lys-48, and Lys-63) and its N-terminal amino group can participate in inter-ubiquitin linkages. Different ubiquitin ligases are specific both for targets (substrates) to be ubiquitinylated and for the lysine side chains on the ubiquitins that participate in the inter-ubiquitin isopeptide linkages (Lys-63 or Lys-48, etc.) (Figure 3-36). These multiple forms of ubiquitinylation result in the generation of a wide variety of recognition surfaces that can participate in many protein-protein interactions with the hundreds of proteins (>200 in humans) that contain more than a dozen distinct ubiquitin-binding domains (UBD). In addition, any given polyubiquitin chain has the potential to bind simultaneously to more than one UBD-containing protein, leading to the formation of ubiquitinylation-dependent multiprotein complexes. Some deubiquitinases can remove an intact polyubiquitin chain from a modified protein (“anchored” chain) and thus generate a polyubiquitin chain not covalently linked to another protein (“unanchored” chain). Even these unanchored chains may serve a regulatory role. With this great structural diversity, it is not surprising that cells use ubiquitinylation and deubiquitinylation to control many different cellular functions.

We have already seen how polyubiquitinylation via Lys-48 residues is used to tag proteins for proteasomal degradation.





**FIGURE 3-36 Determination of polyubiquitin function by the lysine used for inter-ubiquitin isopeptide bonds.** Different ubiquitin ligases catalyze polyubiquitinylation of distinct target (substrate) proteins (colored ovals) using distinct lysine side chains of ubiquitin molecules (purple) to generate the inter-ubiquitin isopeptide linkages (blue) with Gly-76 of the adjacent ubiquitin. Dotted blue arrows represent additional ubiquitins in the chain that are not shown. The lysine used

for the isopeptide bonds determines the function of the polyubiquitinylation. For example, polyubiquitins with Lys-48:Gly-76 isopeptide bonds direct the target to proteasomes for degradation. Those that use Lys-63, Lys-33, and Lys-11 influence signaling, T-lymphocyte control, and cell division, respectively. Isopeptide bonds involving ubiquitin's Lys-6, Lys-27, and Lys-29 and bonds using its N-terminal amino group (not shown) can also be used to generate polyubiquitin chains.

There is evidence that polyubiquitinylation via other Lys residues (for example, Lys-11 and Lys-33, but not Lys-63) can also target proteins for proteasomal degradation. Strikingly, ubiquitinylation unrelated to protein degradation can also control diverse cell functions, including cellular internalization of molecules via endocytosis (see Chapter 14), repair of damaged DNA, metabolism, messenger RNA synthesis (transcription), defense against pathogens, cell division/cell cycle progression, cell signaling pathways, trafficking of proteins within a cell, and apoptosis. The lysine used to form the inter-ubiquitin isopeptide bonds can vary depending on the cellular system that is regulated (see Figure 3-36). For example, polyubiquitinylation with Lys-63 linkages is used in many cellular identification and signaling systems, such as recognition of the presence of intracellular viruses and bacteria and the consequent induction of a protective immune response, as well as direction of these pathogens to lysosomes for degradation. Lys-11-linked polyubiquitin chains regulate cell division. Lys-33-linked chains help suppress the activity of receptors on specialized white blood cells, called T lymphocytes (see Chapter 23), and so control the activity and function of those lymphocytes.

## Proteolytic Cleavage Irreversibly Activates or Inactivates Some Proteins

Unlike phosphorylation and ubiquitinylation, which are reversible, the activation or inactivation of protein function by proteolytic cleavage is an irreversible mechanism for regulating protein activity. For example, many polypeptide hormones, such as insulin, are synthesized as longer precursors, and prior to secretion from cells some of their peptide bonds must be hydrolyzed for them to fold properly. In some cases, a single long precursor *prohormone* polypeptide is cleaved into several distinct active hormones. To prevent the pancreatic serine proteases from inappropriately digesting proteins before they reach the small intestine, they are synthesized as *zymogens*, inactive precursor enzymes. Cleavage of a peptide bond near the N-terminus of trypsinogen (the zymogen of trypsin) by a highly specific protease in the small intestine generates a new N-terminal residue (Ile-16), whose amino group can form an ionic bond with the carboxylic acid side chain of an internal aspartic acid. This binding causes a conformational change that opens the substrate-binding site, activating the enzyme. The active trypsin can then activate

trypsinogen, chymotrypsinogen, and other zymogens. Similar but more elaborate protease cascades (with one protease activating inactive precursors of others) that can amplify an initial signal play important roles in several systems, such as the blood-clotting cascade and the complement system (see Chapter 23). The importance of carefully regulating such systems is clear—inappropriate clotting, for example, could fatally clog the circulatory system, while insufficient clotting could lead to uncontrolled bleeding.

An unusual and rare type of proteolytic processing, termed *protein self-splicing*, takes place in bacteria and some eukaryotes. This process is analogous to editing film: an internal segment of a polypeptide is removed and the ends of the polypeptide are rejoined (ligated). Unlike other forms of proteolytic processing, protein self-splicing is an autocatalytic process, which proceeds by itself without the participation of other enzymes. The excised peptide appears to eliminate itself from the protein by a mechanism similar to that used in the processing of some RNA molecules (see Chapter 10). In vertebrate cells, the processing of some proteins includes self-cleavage, but the subsequent ligation step is absent. One such protein is Hedgehog, a membrane-bound signaling molecule that is critical to a number of developmental processes (see Chapter 16).

### Higher-Order Regulation Includes Control of Protein Location

All the regulatory mechanisms heretofore described affect a protein locally at its site of action, altering the protein's concentration or turning its activity on or off. Normal functioning of a cell, however, also requires the segregation of proteins to particular compartments, such as the mitochondria, nucleus, or lysosomes. In regard to enzymes, compartmentation not only provides an opportunity for controlling the delivery of substrate or the exit of product, but also permits competing reactions to take place simultaneously in different parts of a cell. We describe the mechanisms that cells use to direct various proteins to different compartments in Chapters 13 and 14.

## KEY CONCEPTS OF SECTION 3.4

### Regulating Protein Function

- Proteins may be regulated at the level of protein synthesis, protein degradation, or the intrinsic activity of proteins through noncovalent or covalent interactions.
- The life span of intracellular proteins is largely determined by their susceptibility to proteolytic degradation.
- Many proteins are marked for destruction with a polyubiquitin tag by ubiquitin ligases and then degraded within proteasomes, large cylindrical complexes with multiple protease active sites in their interior chambers (see Figure 3-31).

- Ubiquitylation of proteins is reversible due to the activity of deubiquitylating enzymes.
- In allosteric, the noncovalent binding of one ligand molecule, the allosteric effector, induces a conformational change that alters a protein's activity or affinity for other ligands. The allosteric effector can be identical in structure to or different from the other ligands, whose binding it affects. The allosteric effector can be an activator or an inhibitor.
- In multimeric proteins, such as hemoglobin, that bind multiple identical ligand molecules (e.g., oxygen), the binding of one ligand molecule may increase or decrease the protein's affinity for subsequent ligand molecules. This type of allosteric is known as cooperativity (see Figure 3-32).
- Several allosteric mechanisms act as switches, turning protein activity on and off in a reversible fashion.
- Two classes of intracellular switch proteins regulate a wide variety of cellular processes: (1)  $\text{Ca}^{2+}$ -binding proteins (e.g., calmodulin) and (2) members of the GTPase superfamily (e.g., Ras), which cycle between active GTP-bound and inactive GDP-bound forms (see Figure 3-34). GTPases participate in complex regulatory networks that include proteins (GAP, GEF, GDI) that regulate the cycling of the GTPase between its active and inactive forms.
- The phosphorylation and dephosphorylation of hydroxyl groups on serine, threonine, or tyrosine side chains by protein kinases and phosphatases provide reversible on/off regulation of numerous proteins (see Figure 3-35).
- Variations in the nature of the covalent attachment of ubiquitin to proteins (mono-, multi-, and polyubiquitylation involving a variety of linkages between the ubiquitin monomers) are involved in a wide variety of cellular functions other than proteasome-mediated degradation, such as changes in the location or activity of proteins (see Figure 3-36).
- Many types of covalent and noncovalent regulation are reversible, but some forms of regulation, such as proteolytic cleavage, are irreversible.
- Higher-order regulation includes the intracellular location, or compartmentation, of proteins.

## 3.5 Purifying, Detecting, and Characterizing Proteins

A protein often must be purified before its structure and the mechanism of its action can be studied in detail. However, because proteins vary in size, shape, oligomerization state, charge, and water solubility, no single method can be used to isolate all proteins. To isolate one particular protein from the estimated 10,000 different proteins in a particular type of cell is a daunting task that requires methods both for separating proteins and for detecting the presence of specific proteins.

Any molecule, whether protein, carbohydrate, or nucleic acid, can be separated, or *resolved*, from other molecules on the basis of their differences in one or more physical or chemical characteristics. The larger and more numerous the differences between two proteins, the easier and more efficient their separation. As a practical matter, the more abundant a particular protein is in a biological sample, the easier it is to separate it from the other molecules in the sample. The three most widely used characteristics for separating proteins are *size*, defined as either length or mass; net electrical charge; and *affinity* for specific ligands. In this section, we briefly outline several important techniques for separating proteins; these separation techniques are also useful for the separation of nucleic acids and other biomolecules. (Specialized methods for removing membrane proteins from membranes are described in Chapter 7 after the unique properties of these proteins are discussed.) We then consider the use of radioactive compounds for tracking biological activity. Finally, we consider several techniques for characterizing a protein's mass, sequence, and three-dimensional structure.

## Centrifugation Can Separate Particles and Molecules That Differ in Mass or Density

The first step in a typical protein purification scheme is centrifugation. The principle behind centrifugation is that two types of particles in suspension (cells, cell fragments, organelles, or molecules) with different masses or densities will settle to the bottom of a test tube at different rates. Remember, mass is the weight of a sample (measured in daltons or molecular weight units), whereas density is the ratio of its mass to volume (often expressed as grams per liter because of the methods used to measure density). Proteins vary greatly in mass, but not in density. Unless a protein has an attached lipid or carbohydrate, its density will not vary by more than 15 percent from  $1.37 \text{ g/cm}^3$ , the average protein density. Heavier or denser molecules settle, or *sediment*, more quickly than lighter or less dense molecules.

A centrifuge speeds sedimentation by subjecting particles in suspension to centrifugal forces as great as 1 million times the force of gravity,  $g$ , which can sediment particles as small as 10 kDa. Modern ultracentrifuges achieve these forces by reaching speeds of 150,000 revolutions per minute (rpm) or greater. However, small particles with masses of 5 kDa or less will not sediment uniformly even at such remarkably high rotation rates. The extraordinary technical achievements of modern ultracentrifuges can be appreciated by considering that they can rotate a several-pound rotor (about the size of an American football) that holds the samples in tubes at rates as high as 2500 revolutions per second!

Centrifugation is used for two basic purposes: (1) as a preparative technique to separate one type of material from others with the goal of obtaining enough of the material to perform subsequent experiments and (2) as an analytical technique to measure physical properties (e.g., molecular

weight, density, shape, and equilibrium binding constants) of macromolecules. The sedimentation constant,  $s$ , of a protein is a measure of its sedimentation rate. The sedimentation constant is commonly expressed in Svedberg units (S), where a typical large protein complex is about 3–5S, a proteasome is 26S, and a eukaryotic ribosome is 80S.

**Differential Centrifugation** The most common initial step in protein purification from cells or tissues is the separation of water-soluble proteins from insoluble cellular material by *differential centrifugation*. A starting mixture, commonly a cell homogenate (mechanically broken cells), is poured into a tube and spun at a rotor speed, and for a period of time, that forces cell organelles such as nuclei as well as large unbroken cells or large cell fragments to collect as a pellet at the bottom; the soluble proteins remain in the supernatant (Figure 3-37a). The supernatant fraction then is poured off, and either it or the pellet can be subjected to other purification methods to separate the many different proteins that they contain.

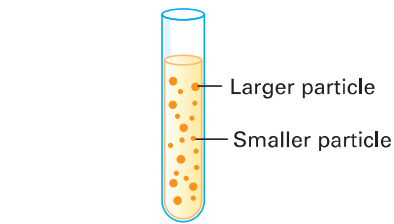
**Rate-Zonal Centrifugation** On the basis of differences in their masses, water-soluble proteins can be separated by centrifugation through a solution of increasing density, called a *density gradient*. A concentrated sucrose solution is commonly used to form a density gradient in a centrifuge tube (with higher concentrations of sucrose, and thus a higher solution density, toward the bottom of the tube, Figure 3-37b). When a protein mixture is placed on top of a sucrose density gradient in a tube and subjected to centrifugation, each protein in the mixture migrates down the tube at a rate controlled by the protein's physical properties. All the proteins start from the thin layer of the sample that was placed at the top of the tube and separate into bands (actually, disks) of proteins of different masses as they travel at different rates through the gradient. In this separation technique, called *rate-zonal centrifugation*, samples are centrifuged just long enough to separate the molecules of interest into discrete bands, also called zones (see Figure 3-37b). If a sample is centrifuged for too short a time, the different protein molecules will not separate sufficiently. If a sample is centrifuged much longer than necessary, all the proteins will end up mixed together at the bottom of the tube.

Although the sedimentation rate is strongly influenced by particle mass, rate-zonal centrifugation is seldom effective in determining precise molecular weights because variations in shape also affect the sedimentation rate. The exact effects of shape are hard to assess, especially for proteins or other molecules, such as single-stranded nucleic acid molecules, that can assume many complex shapes. Nevertheless, rate-zonal centrifugation has proved to be a practical method for separating many different types of polymers and particles. A second density-gradient technique, called *equilibrium density-gradient centrifugation*, is used mainly to separate DNA, lipoproteins that carry lipids through the circulatory system, or organelles (see Figure 4-37).

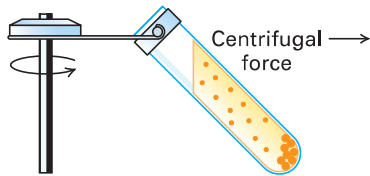


(a) Differential centrifugation

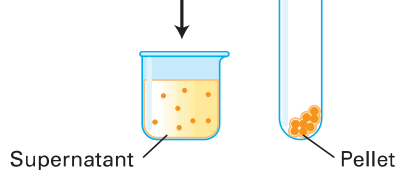
**1** Sample is poured into tube



**2** Centrifuge  
Particles settle  
according to  
mass

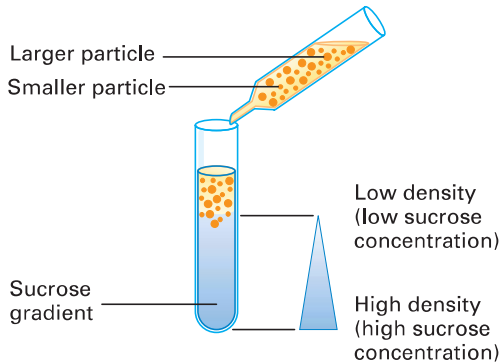


**3** Stop centrifuge  
Decant liquid  
into container

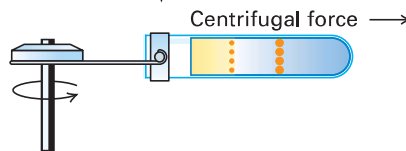


(b) Rate-zonal centrifugation

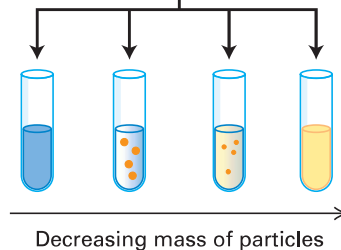
**1** Sample is layered on top of density gradient



**2** Centrifuge  
Particles settle  
according to  
mass



**3** Stop centrifuge  
Collect fractions  
and do assay



**EXPERIMENTAL FIGURE 3-37**  
**Centrifugation techniques separate particles that differ in mass or density.** (a) In differential centrifugation, a cell homogenate or other mixture is spun long enough to sediment the larger particles (e.g., cell organelles, cells), which collect as a pellet at the bottom of the tube (step **2**). The smaller particles (e.g., soluble proteins, nucleic acids) remain in the liquid supernatant, which can be transferred to another tube (step **3**). (b) In rate-zonal centrifugation, a mixture is spun (step **1**) just long enough to separate molecules that differ in mass but may be similar in shape and density (e.g., globular proteins, RNA molecules) into discrete zones within a density gradient commonly formed by a concentrated sucrose solution. Fractions are removed from the bottom of the tube and subjected to testing (assayed).

## Electrophoresis Separates Molecules on the Basis of Their Charge-to-Mass Ratio

Electrophoresis, a technique for separating molecules in a mixture under the influence of an applied electric field, is one of the most frequently used techniques to study proteins and nucleic acids. Dissolved molecules in an electric field move, or migrate, at a speed determined by their charge-to-mass (charge:mass) ratio and the physical properties of the medium through which they migrate. For example, if two molecules have the same mass and shape, the one with the greater net charge will move faster toward an electrode of the opposite polarity.

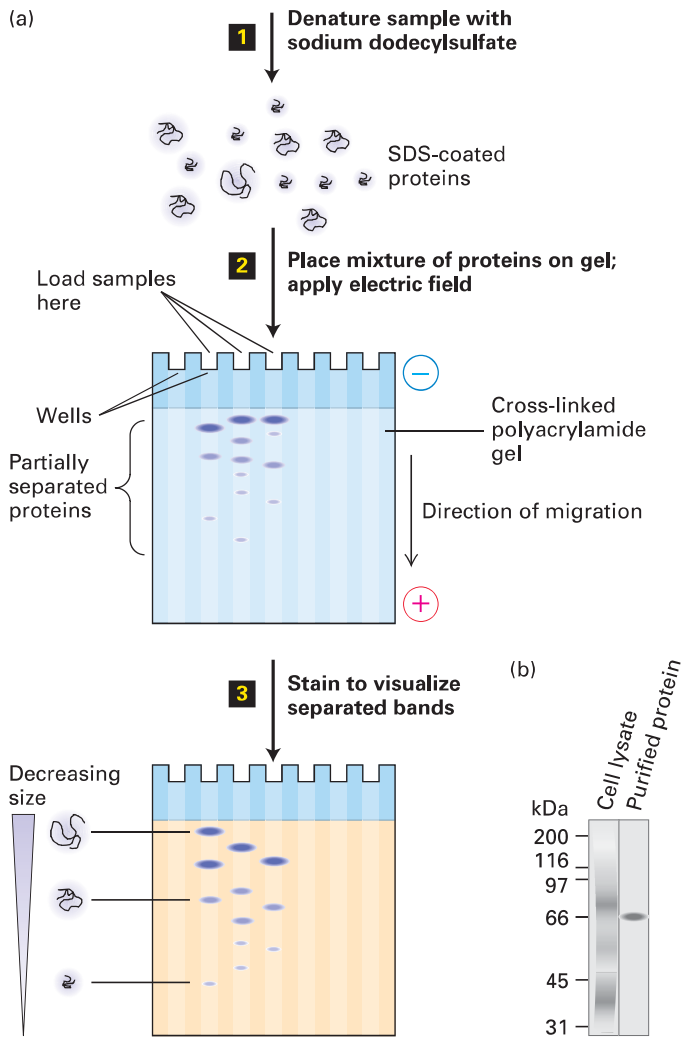
**SDS-Polyacrylamide Gel Electrophoresis** Because many proteins or nucleic acids that differ in size and shape have nearly identical charge:mass ratios, electrophoresis of these macromolecules in a liquid solution results in little or no separation of molecules of different lengths. However, successful separation of proteins and nucleic acids can be

accomplished by electrophoresis in various gels (semisolid suspensions in water similar to the congealed gelatin found in desserts). These gels are commonly cast into flat, relatively thin slabs between a pair of glass plates. When a mixture of proteins is placed in a gel and an electric current is applied, the gel acts as a sieve, allowing smaller species to maneuver more rapidly through its pores than larger species do. The shape of a molecule can also influence its rate of migration (long asymmetric molecules migrate more slowly than spherical ones of the same mass).

Electrophoretic separation of proteins is most commonly performed in polyacrylamide gels. These gels are made by polymerizing a solution of acrylamide monomers into polyacrylamide chains and simultaneously cross-linking the chains into a semisolid matrix. The pore size of a gel can be varied by adjusting the concentrations of polyacrylamide and the cross-linking reagent. The rate at which a protein moves through a gel is influenced by the gel's pore size and the strength of the electric field. By suitable adjustment of

these parameters, proteins of widely varying sizes can be resolved (separated from one another) by this technique, known as *polyacrylamide gel electrophoresis* (PAGE).

In the most powerful technique for resolving protein mixtures, proteins are exposed to the ionic detergent SDS (sodium dodecylsulfate) before and during gel electrophoresis (Figure 3-38). SDS denatures proteins, in part because

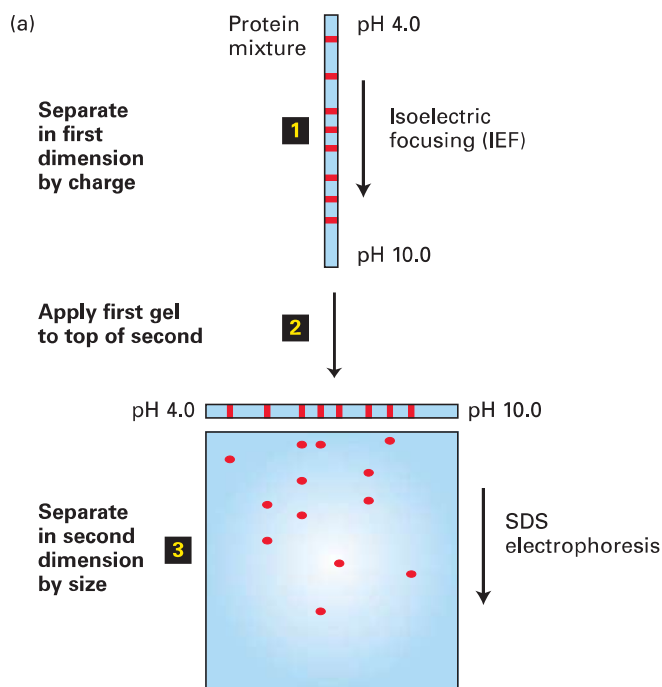


**EXPERIMENTAL FIGURE 3-38 SDS-polyacrylamide gel electrophoresis (SDS-PAGE) separates proteins primarily on the basis of their masses.** (a) Initial treatment with SDS, a negatively charged detergent, dissociates multimeric proteins and denatures all the polypeptide chains (step 1). During electrophoresis, the SDS-protein complexes migrate through the polyacrylamide gel (step 2). Small complexes are able to move through the pores faster than larger ones. Thus the proteins separate into bands according to their sizes as they migrate. The separated protein bands are visualized by staining with a dye (step 3). (b) Example of SDS-PAGE separation of all the proteins in a whole-cell lysate (detergent-solubilized cells). (Left) The many separate stained proteins appear almost as a continuum. (Right) A single protein purified from the lysate by a single step of antibody-affinity chromatography. The proteins were visualized by staining with a silver-based dye. [Part (b) data from B. Liu and M. Krieger, 2002, *J. Biol. Chem.* 277:34125–34135]

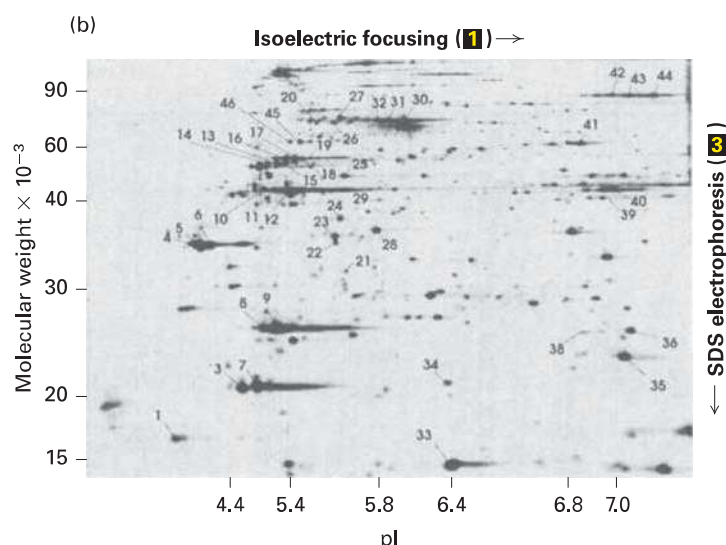
it binds to hydrophobic side chains, destabilizing the hydrophobic interactions in the core of a protein that contribute to its stable conformation. SDS treatment is usually combined with heating, often in the presence of reducing agents that break disulfide bonds. Under these conditions, most multimeric proteins dissociate into their subunits. Typically, the amount of SDS that binds to the protein is proportional to the length of the polypeptide chain and relatively independent of the sequence. Two proteins of similar size will bind the same absolute quantity of SDS, whereas a protein twice that size will bind twice the amount of SDS. Denaturation of a complex protein mixture with SDS in combination with heat usually forces each polypeptide chain into an extended conformation and imparts on each of the proteins in the mixture a constant charge:mass ratio because the dodecylsulfate, which is negatively charged, is the major contributor of charge. As the SDS-bound proteins move through the polyacrylamide gel, they are separated according to size by the sieving action of the gel. SDS treatment thus eliminates the effect of differences in native conformation; therefore, chain length, which is proportional to mass, is the principal determinant of the migration rate of proteins in *SDS-polyacrylamide electrophoresis* (SDS-PAGE). Even chains that differ in molecular weight by less than 10 percent can be resolved by this technique. Moreover, the molecular weight of a protein can be estimated by comparing the distance that it migrates through a gel with the distances that proteins of known molecular weight (called molecular weight “standards”) migrate in the same gel (there is a roughly linear relationship between migration distance and the log of the molecular weight). Proteins within the gels can be extracted for further analysis (e.g., identification by the methods described below).

If two or more polypeptides are cross-linked by disulfide bonds, the protein’s migration rate in SDS-PAGE will depend on whether or not the protein has been reduced to break those bonds prior to electrophoresis. The cross-linked proteins will appear larger than the individual, reduced subunits. By examining samples with and without reduction, one can identify such proteins and their component polypeptides.

**Two-Dimensional Gel Electrophoresis** Electrophoresis of a mixture containing all cellular proteins by SDS-PAGE can separate proteins having relatively large differences in mass, but cannot readily resolve proteins having similar masses (e.g., a 41-kDa protein versus a 42-kDa protein). To separate proteins of similar masses, another physical characteristic must be exploited. Most commonly, this characteristic is electric charge, which is determined by the pH of the sample and by the relative number of the protein’s positively and negatively charged groups, which is in turn dependent on the  $pK_a$ ’s of the ionizable groups (see Chapter 2) on the proteins (usually the amino and carboxyl termini and side chains such as those in lysine and aspartic acid). Two unrelated proteins having similar masses are unlikely to have identical net charges because their sequences, and thus the number of acidic and basic residues, are different.



**EXPERIMENTAL FIGURE 3-39 Two-dimensional gel electrophoresis separates proteins on the basis of charge and mass.** (a) In this technique, proteins are first separated into bands on the basis of their charges by isoelectric focusing (step 1). The resulting gel strip is applied to an SDS-polyacrylamide gel (step 2), and the proteins are separated into spots by mass (step 3). (b) In this two-dimensional



electrophoresis gel of a protein extract from cultured cells, each spot represents a single polypeptide. Polypeptides can be detected by dyes, as here, or by other techniques, such as autoradiography. Each polypeptide is characterized by its isoelectric point (pI) and molecular weight. [Part (b) Michael J. Dunn.]

In *two-dimensional gel electrophoresis*, proteins are separated sequentially, first by their charges and then by their masses (Figure 3-39a). In the first step, a cell or tissue extract is fully denatured by high concentrations (8 M) of urea (and sometimes SDS) and then layered on a strip of gel that contains urea, which removes any bound SDS, and a continuous pH gradient. The pH gradient is formed by ampholytes, polyanionic and polycationic small molecules that are cast into the gel. When an electric field is applied to the gel, the ampholytes will migrate. Ampholytes with an excess of negative charges will migrate toward the anode, where they establish an acidic pH (many protons), while ampholytes with an excess of positive charges will migrate toward the cathode, where they establish an alkaline pH. The careful choice of the mixture of ampholytes and careful preparation of the gel allows the construction of stable pH gradients ranging from pH 3 to pH 10. A charged protein placed at one end of such a gel will migrate through the gradient under the influence of the electric field until it reaches its **isoelectric point (pI)**, the pH at which the net charge of the protein is zero. With no net charge, the protein will migrate no further. This technique, called *isoelectric focusing (IEF)*, can resolve proteins that differ by only one charge unit. This method is sensitive enough to separate phosphorylated and nonphosphorylated versions of the same protein.

Proteins that have been separated on an IEF gel can then be separated in a second dimension on the basis of

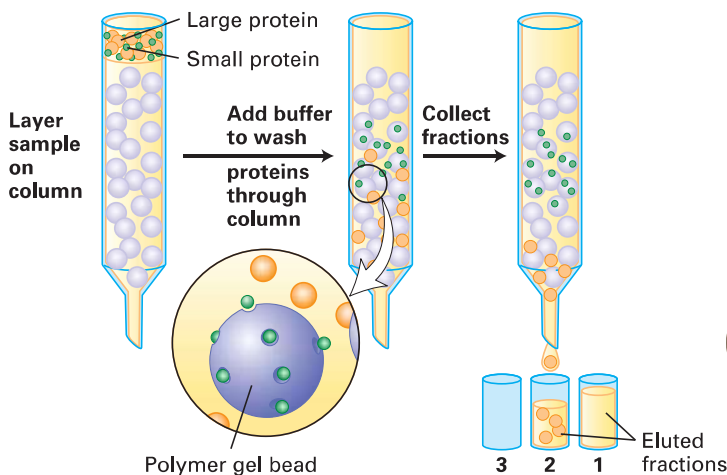
their molecular weights. To accomplish this separation, the IEF gel strip is placed lengthwise on one outside edge of a square or rectangular slab of polyacrylamide gel, this time saturated with SDS to confer on each separated protein a more or less constant charge:mass ratio. When an electric field is imposed, the proteins will migrate from the IEF gel into the SDS gel and then separate according to their masses. The sequential resolution of proteins by charge and mass can achieve excellent separation of cellular proteins and provides a powerful visual representation of the complexity of proteins in cells (Figure 3-39b). Today sophisticated mass spectrometry methods, described below, are often used in place of two-dimensional gel electrophoresis, both to separate and to identify the protein components of a complex sample as well as to compare changes in the amounts of those components in different biological specimens.

### Liquid Chromatography Resolves Proteins by Mass, Charge, or Affinity

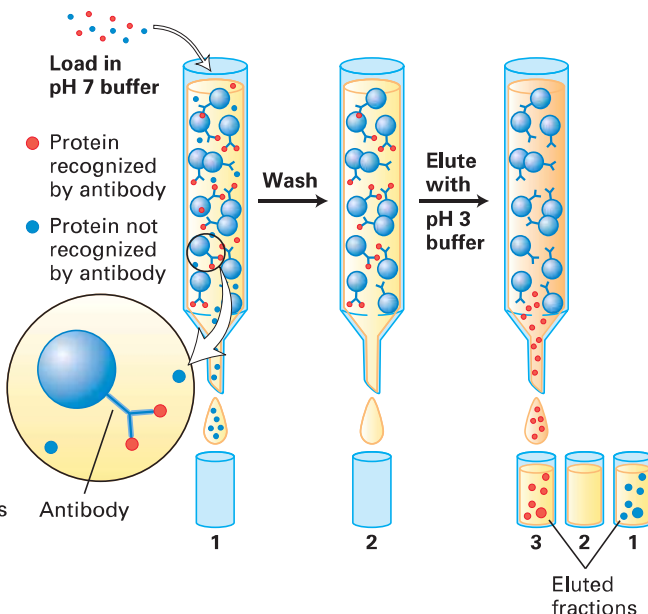
A third common technique for separating mixtures of proteins or fragments of proteins, as well as other molecules, is based on the principle that molecules in solution can differentially interact with (bind to and dissociate from) a particular solid surface, depending on the physical and chemical properties of the molecule and the surface. If the solution is allowed to flow across the surface, then molecules that



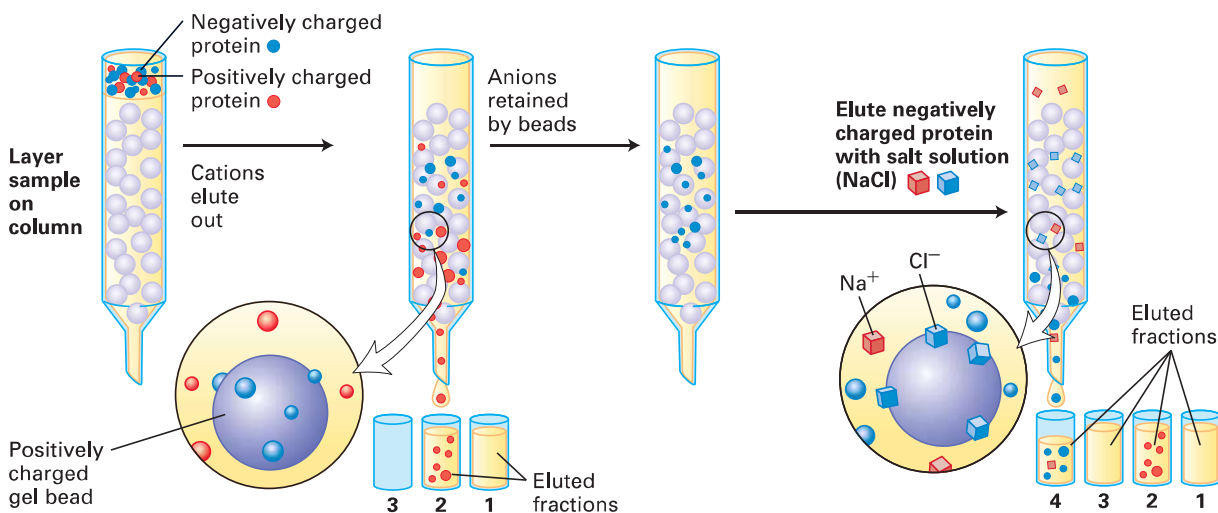
(a) Gel filtration chromatography



(c) Antibody-affinity chromatography



(b) Ion-exchange chromatography



**EXPERIMENTAL FIGURE 3-40 Three commonly used liquid chromatographic techniques separate proteins on the basis of mass, charge, or affinity for a specific binding partner.** (a) Gel filtration chromatography separates proteins that differ in size. A mixture of proteins is carefully placed, or loaded, on the top of a cylinder packed with porous beads. Smaller proteins travel through the column more slowly than larger proteins. Thus the different proteins emerging in the eluate flowing out of the bottom of the column at different times (different elution volumes) can be collected in separate tubes, called fractions. (b) Ion-exchange chromatography separates proteins that differ in net charge in columns packed with beads that carry either a positive charge (shown here) or a negative charge. Proteins having the same net charge as the beads are repelled and flow through the column,

whereas proteins having the opposite charge bind to the beads more or less tightly, depending on their structures. Bound proteins—in this case, negatively charged proteins—are subsequently eluted by passing a salt gradient (usually of NaCl or KCl) through the column. As the ions bind to the beads, they displace the proteins; more tightly bound proteins require higher salt concentrations in order to be released. (c) In antibody-affinity chromatography, a mixture of proteins is passed through a column packed with beads to which a specific antibody is covalently attached. Only proteins with high affinity for the antibody are retained by the column; all the nonbinding proteins flow through. After the column is washed, the bound protein is eluted with an acidic solution or some other solution that disrupts the antigen-antibody complexes; the released protein then flows out of the column and is collected.

interact frequently with the surface will spend more time bound to the surface, and thus flow past the surface more slowly, than molecules that interact infrequently with it. In this technique, called *liquid chromatography (LC)*, the sample is placed on top of a tightly packed column of spherical beads held within a glass, metal, or plastic cylinder

(Figure 3-40). The sample then flows down the column, driven by gravitational or hydrostatic forces alone or sometimes with the assistance of a pump. In some LC systems, the composition of the fluid flowing out of the column is monitored continuously (for example, by spectroscopy). Small aliquots of fluid flowing out of the column, called *fractions*,

are collected sequentially and can be analyzed subsequently for their contents and chemical activities (e.g., enzymatic activity). The nature of the beads in the column determines whether the separation of proteins depends on differences in their mass, charge, or other binding properties (e.g., affinity for substances attached to the beads).

**Gel Filtration Chromatography** Proteins that differ in mass can be separated on a column of porous beads made from polyacrylamide, dextran (a bacterial polysaccharide), or agarose (a seaweed derivative)—a technique called gel filtration chromatography. Although proteins flow around the beads, they spend some time within the large depressions that cover a bead's surface. Because smaller proteins can penetrate these depressions more readily than larger proteins can, they travel through a gel filtration column more slowly than larger proteins do (Figure 3-40a). (In contrast, proteins migrate *through* the pores in an electrophoretic gel; thus smaller proteins move faster than larger ones.) The total volume of liquid required to elute (or separate and remove) a protein from a gel filtration column depends on the protein's mass: the smaller its mass, the longer it is trapped on the beads, the longer it takes to traverse the column, and the greater the elution volume. If proteins of known mass are used as standards to calibrate the column, the elution volume can be used to estimate the mass of a protein in a mixture. A protein's shape as well as its mass can influence the elution volume.

**Ion-Exchange Chromatography** In ion-exchange chromatography, proteins are separated on the basis of differences in their charges. This technique makes use of specially modified beads whose surfaces are covered by amino groups or carboxyl groups and thus carry either a positive charge ( $\text{NH}_3^+$ ) or a negative charge ( $\text{COO}^-$ ) at neutral pH.

The proteins in a mixture carry various net charges at any given pH. When a solution of mixed proteins flows through a column of positively charged beads, only proteins with a net negative charge (acidic proteins) adhere to the beads; neutral and positively charged (basic) proteins flow unimpeded through the column (Figure 3-40b). The acidic proteins are then eluted selectively from the column by passing a solution of increasing concentrations of salt (a salt gradient) through the column. At low salt concentrations, protein molecules and beads are attracted by their opposite charges. At higher salt concentrations, negatively charged salt ions bind to the positively charged beads, displacing the negatively charged proteins. In a gradient of increasing salt concentrations, weakly bound proteins—those with a relatively low charge—are eluted first, and highly charged proteins are eluted last. Similarly, a negatively charged column can be used to retain and fractionate basic (positively charged) proteins.

**Affinity Chromatography** The ability of proteins to bind specifically to other molecules is the basis of affinity chromatography. In this technique, ligands or other molecules that bind to the protein of interest are covalently attached to the beads used to form the column. Ligands can be enzyme substrates, inhibitors or their analogs, or other small molecules that bind

to specific proteins. In a widely used form of this technique—*antibody-affinity*, or *immunoaffinity chromatography*—the molecule attached to the beads is an antibody specific for the desired protein (Figure 3-40c). (We discuss antibodies as tools for studying proteins next; see also Chapter 23, which describes how antibodies are made.)

In principle, an affinity column will retain only those proteins that bind the molecule attached to the beads; the remaining proteins, regardless of their charges or masses, will pass through the column because they do not bind. However, if a retained protein is in turn bound to other molecules, forming a complex, then the entire complex is retained on the column. The proteins bound to the affinity column are then eluted by adding an excess of a soluble form of the ligand, by exposure of bound materials to detergents, or by changing the salt concentration or pH such that the binding to the molecule on the column is disrupted. The ability of this technique to separate particular proteins depends on the selection of appropriate binding partners that bind more tightly to the protein of interest than to other proteins.

### Highly Specific Enzyme and Antibody Assays Can Detect Individual Proteins

The purification of a protein, or any other molecule, requires a specific *assay* that can detect the presence of that molecule as it is separated from other molecules (e.g., in column or density-gradient fractions or gel bands or spots). Such an assay capitalizes on some highly distinctive characteristic of a protein: the ability to bind a particular ligand, to catalyze a particular reaction, or to be recognized by a specific antibody. The assay must also be simple and fast to minimize errors and the possibility that the protein of interest will become denatured or degraded while the assay is being performed. The goal of any purification scheme is to isolate sufficient amounts of a given protein for study; thus a useful assay must also be sensitive enough that only a small proportion of the available material is consumed by it. Many common protein assays require just  $10^{-9}$  to  $10^{-12}$  g of material.

**Chromogenic Enzyme Reactions** Many assays are tailored to detect some functional aspect of a protein. For example, assays of enzymatic activity are based on the ability to detect the loss of substrate or the formation of product. Some enzymatic activity assays use chromogenic substrates, which change color in the course of the reaction. (Some substrates are naturally chromogenic; those that are not can be linked to a chromogenic molecule.) Because of the specificity of an enzyme for its substrate, only samples that contain the enzyme will change color in the presence of a chromogenic substrate; the rate of the change provides a measure of the quantity of enzyme present. Enzymes that catalyze chromogenic reactions can also be fused or chemically linked to an antibody and used to “report” the presence or location of an antigen to which the antibody binds (see below).

**Antibody Assays** As noted earlier, antibodies have the distinctive characteristic of binding tightly and specifically

to antigens. As a consequence, preparations of antibodies that recognize a protein antigen of interest can be generated and used to detect the presence of that protein, either in a complex mixture of other proteins (finding a needle in a haystack, as it were) or in a partially or completely purified preparation of a particular protein. The presence of the antigen can be detected by labeling the antibody with an enzyme, a fluorescent molecule, or a radioactive isotope, which can be detected using an enzyme assay, fluorescence microscopy or spectroscopy, or a radiation detector, respectively. For example, luciferase, an enzyme present in fireflies and some bacteria, can be linked to an antibody. In the presence of ATP and its substrate, luciferin, luciferase catalyzes a light-emitting reaction. In either case, after the antibody binds to the protein of interest (the antigen) and unbound antibody is washed away, substrates of the linked enzyme are added and the appearance of color or emitted light is monitored. The intensity is proportional to the amount of enzyme-linked antibody, and thus antigen, in the sample. Alternatively, after a first (or “primary”) antibody that is not otherwise labeled binds to its target protein, a second (“secondary”), labeled antibody that can recognize the first antibody is used to bind to the complex of the first antibody and its target. This combination of primary and secondary antibodies (sometimes called an antibody “sandwich”) permits very high sensitivity in the detection of a target protein because the labeled secondary antibody is often a mixture of antibodies that bind to multiple sites on the first antibody and thus results in a stronger signal than labeling of the primary antibody alone. It is important to remember that an antibody recognizes and binds to only its epitope on a target antigen. If that epitope is altered—for example, by partial unfolding or post-translational modifications—or is blocked when the antigen protein is bound to some other molecule, the ability of the antibody to bind may be reduced or completely lost. Thus the absence of antibody binding does not necessarily mean that the antigen is not present in a sample, only that the epitope portion of that antigen is not present or accessible for antibody binding.

To generate antibodies to a protein (discussed in detail in Chapter 23), the intact protein, or a fragment of the protein, is injected into an animal (usually a rabbit, mouse, or goat). Sometimes a short synthetic peptide of 10–15 residues based on the sequence of the protein of interest is used as the antigen to induce antibody formation. Such a synthetic peptide, when coupled to a large protein carrier, can induce an animal to produce antibodies that bind specifically to that part (the epitope) of the full-sized, natural protein. Biosynthetically or chemically attaching the epitope to an unrelated protein is called *epitope tagging*. As we’ll see throughout this book, antibodies generated using either synthetic peptide epitopes or intact proteins are extremely versatile reagents for isolating, detecting, and characterizing proteins.

**Detecting Proteins by Attaching Green Fluorescent Protein** An alternative to epitope tagging that is particularly useful in detecting specific proteins within live cells makes use of *green fluorescent protein (GFP)*, a naturally fluorescent

protein found in jellyfish (see Figure 4-16). A chimeric protein containing both the protein of interest and GFP, linked together in one polypeptide chain, is expressed in cells by introducing into the cells a gene encoding the combined protein. The amounts and intracellular distribution of the chimeric protein can then be determined readily. This chimeric protein approach is described in Chapter 4.

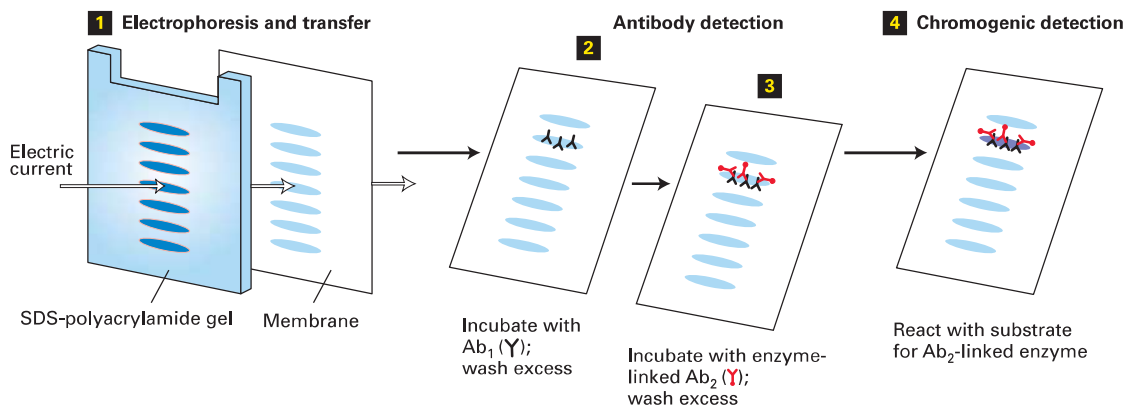
**Detecting Proteins in Gels** Proteins embedded within a gel usually are not visible. The two general approaches to detecting proteins in gels are either to label or stain the proteins while they are still within the gel or to electrophoretically transfer the proteins to a membrane made of nitrocellulose or polyvinylidene difluoride and then detect them. Proteins within gels are usually stained with an organic dye or a silver-based stain, both of which are detectable with normal visible light, or with a fluorescent dye, which requires specialized detection equipment. Coomassie blue, the most commonly used organic dye, is typically used to detect about 1000 ng of protein, with a lower limit of detection of about 4–10 ng. Silver staining and fluorescence staining are more sensitive (with a lower limit of ~1 ng). Coomassie and other stains can also be used to visualize proteins after transfer to membranes; however, the most common method of visualizing proteins in membranes is immunoblotting.

**Immunoblotting**, also called *Western blotting*, combines the resolving power of gel electrophoresis with the specificity of antibodies. This multistep procedure is commonly used to separate proteins and then identify a specific protein of interest. As shown in Figure 3-41, two different antibodies are used, one that is specific for the protein of interest (primary antibody) and a secondary antibody that binds to the first and is linked to an enzyme or other molecule that permits detection of the first antibody (and thus the protein of interest to which it binds). The enzyme to which the second antibody is attached can either generate a visible colored product or, by a process called *chemiluminescence*, produce light that can readily be recorded by film or a sensitive detector. An example of the results of an immunoblotting experiment can be seen in Figure 15-10. If an antibody to the protein of interest is not available, but the gene encoding the protein is available and can be used to express the protein, recombinant DNA methods (see Chapter 5) can incorporate a small peptide epitope into the normal sequence of the protein (epitope tagging) that can be detected by a commercially available antibody to that epitope.

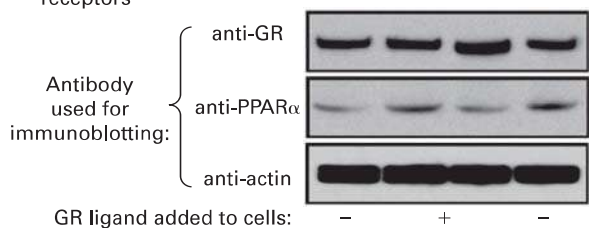
**Immunoprecipitation** Immunoprecipitation, often abbreviated as *IP*, exploits the specificity of antibodies to separate a protein of interest from other molecules in a complex mixture—for example, all proteins extracted from a sample of cells or a sample of blood. An antibody to the protein of interest is added to a sample, and the antibody is given time to bind to epitopes on the target protein. An agent that binds to the antibody is then added to cause the antibody and its bound target to precipitate out of solution into particles that can be isolated by centrifugation. A detailed example of this technique is described in Chapter 15 (p. 684). The precipitate is then solubilized under denaturing conditions—for example, in



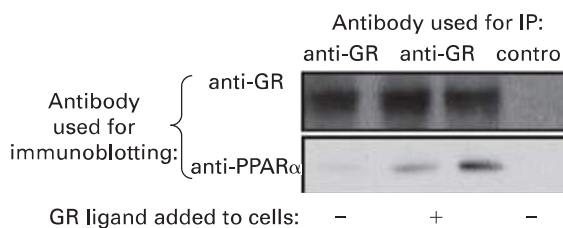
(a) General method of immunoblotting



(b) Immunoblotting of lysed cells to detect intracellular receptors



(c) Immunoprecipitation (IP) followed by immunoblotting (co-IP)



**EXPERIMENTAL FIGURE 3-41 Immunoblotting (IP, or Western blotting) and co-immunoprecipitation (co-IP) can detect specific proteins and their binding partners.** (a) Immunoblotting method. Step **1**:

After a protein mixture has been electrophoresed through an SDS gel, the separated bands (or spots, for two-dimensional gel electrophoresis) are transferred (blotted) from the gel onto a porous membrane from which the protein is not readily removed. Individual proteins (represented by blue ovals) are not visible at this stage. Step **2**: The membrane is flooded with a solution of an antibody ( $Ab_1$ ) specific for the protein of interest and allowed to incubate for a while.  $Ab_1$  binds to the protein of interest (second from the top), but not to any other proteins attached to the membrane, forming a layer of antibody molecules coincident with the protein (whose position still cannot be seen at this point). Then the membrane is washed to remove unbound  $Ab_1$ . Step **3**: The membrane is incubated with a second antibody ( $Ab_2$ ) that specifically recognizes and binds to the first ( $Ab_1$ ). This second antibody is covalently linked to an enzyme that catalyzes a chromogenic reaction or releases light (e.g., chemiluminescence), a radioactive isotope, or some other substance whose presence can be detected with great sensitivity. Step **4**: Finally, the location and amount of bound  $Ab_2$  are detected (e.g., by its color for a chromogenic reaction or by detectors or film that measure the light released by chemiluminescence), permitting the electrophoretic mobility (and therefore the mass) of the protein of interest to be determined as well as its quantity (based on band intensity). (b) Immunoblotting was used to detect intracellular receptors and the influence of exposure to a ligand for one of the receptors. In this experiment, cells that are precursors to red blood cells were maintained *in vitro* in petri dishes and then treated with no ligand (–, leftmost and rightmost lanes) or a ligand that binds to GR, the glucocorticoid receptor (+, center lane). The cells were then lysed in detergent, and immunoblotting (Western blotting) was performed on the total cell lysates using three different antibodies that bind to GR (anti-GR), to a receptor called PPAR $\alpha$  (anti-PPAR $\alpha$ ), or to an abundant intracellular protein, actin, whose presence and abundance was not expected to be sensitive to treatment with the ligand.

The equal intensities of the immunoblotting bands detected using the anti-actin antibody (bottom box) provided a “loading control,” which established that essentially equal amounts of cell lysate were applied (loaded) in each lane of the gel. The approximately equal intensities of the bands for both GR and PPAR $\alpha$  with or without prior incubation of the cells with the GR ligand showed that the ligand did not substantially alter the amounts of either of these proteins in the cells. Portions of the same cell lysates used for the immunoblotting in part (b) were also used for the immunoprecipitation/immunoblotting shown in part (c). (c) Immunoprecipitation (IP) followed by immunoblotting (together called co-IP) was used to determine if the GR ligand can induce formation of a stable complex that contains both GR and PPAR $\alpha$ . Portions of the cell lysates were immunoprecipitated with an antibody to GR (left and center lanes) or a control antibody (right lane) that cannot bind to either GR or PPAR $\alpha$ . The immunoprecipitates were separated from the rest of the lysates by centrifugation and then analyzed by immunoblotting with either anti-GR (top box) or anti-PPAR $\alpha$  (bottom box). As expected, the top box shows that the GR protein was detected in the immunoprecipitates generated using the  $\alpha$ -GR when the same anti-GR antibody was used for the immunoblotting, but not in the immunoprecipitates generated with the control antibody (no band observed). Strikingly, when one examines the immunoprecipitates by immunoblotting with the anti-PPAR $\alpha$  antibody (bottom box), a substantial amount of PPAR $\alpha$  is seen when the GR ligand is present (center lane), whereas little co-precipitates in the absence of the GR ligand (left lane) or in the control immunoprecipitate (right lane). These results indicate that the GR ligand induces formation of a complex containing both the glucocorticoid receptor and the PPAR $\alpha$  proteins. These results do not establish whether or not the GR and PPAR $\alpha$  proteins bind directly to each other when the GR ligand is present or if there are additional molecules in the complex that act as intermediates holding the GR and PPAR $\alpha$  tightly together when the ligand is present. [Parts (b) and (c) reprinted by permission from Macmillan Publishers Ltd, from Lee, H.Y. et al., “PPAR- $\alpha$  and glucocorticoid receptor synergize to promote erythroid progenitor self-renewal,” *Nature*, 2015, **522**:474–477]

a detergent-containing buffer—to separate the antibody from the protein, and the immunoprecipitated target protein can then be analyzed. If the immunoprecipitated target is tightly bound to one or more other molecules, those bound molecules may be precipitated along with the protein of interest (*co-immunoprecipitation*, sometimes abbreviated as *co-IP*). The co-IP method is used frequently to identify and characterize quaternary structures and supramolecular complexes.

## Radioisotopes Are Indispensable Tools for Detecting Biological Molecules

A sensitive method for tracking a protein or other biological molecule is by detecting the radioactivity emitted from a radiolabel introduced into the molecule. In a radiolabeled molecule, at least one atom is present in a radioactive form, called a **radioisotope**.

**Radioisotopes Useful in Biological Research** Hundreds of biological molecules (e.g., amino acids, nucleosides, and numerous other small-molecule metabolites) labeled with various radioisotopes are commercially available. These preparations vary considerably in their *specific activity*, which is the amount of radioactivity per unit of material, measured in disintegrations per minute (dpm) per millimole (mmol). The specific activity of a labeled compound depends on the radioisotope's *half-life*, the time required for half the atoms to undergo radioactive decay, which releases the detectable radiation. In general, the shorter the half-life of a radioisotope, the higher its specific activity (Table 3-1). The specific activity of a labeled compound must be high enough for accurate detection of its emitted radiation.

A common approach to radiolabeling macromolecules (proteins, RNA, DNA) in cells is to add a radiolabeled biosynthetic precursor to the extracellular medium [e.g.,  $^3\text{H}$ - or  $^{35}\text{S}$ -labeled amino acids,  $^{32}\text{P}$ -labeled phosphate (precursor for  $^{32}\text{P}$ -labeled ATP), or  $^3\text{H}$ -labeled nucleic acid precursors such as deoxythymidine (also simply called thymidine) or  $^{32}\text{P}$ -labeled phosphate]. The precursor enters the cells via transporters (see Chapter 11) and is incorporated into newly synthesized macromolecules by the cells (see Chapter 5). For example, methionine and cysteine labeled

with sulfur-35 ( $^{35}\text{S}$ ) are widely used to label cellular proteins because preparations of these amino acids with high specific activities ( $>10^{15}$  dpm/mmol) are available. Kinases within cells (or used *in vitro*) can transfer a  $^{32}\text{P}$ -labeled phosphate from  $^{32}\text{P}$ -labeled ATP to label phosphoproteins. Likewise, commercial preparations of  $^3\text{H}$ -labeled nucleic acid precursors have much higher specific activities than those of the corresponding  $^{14}\text{C}$ -labeled preparations. In most experiments, the former are preferable because they allow RNA or DNA to be adequately labeled a shorter time after incorporation or require a smaller cell sample. Various phosphate-containing compounds in which the phosphorus atom is the radioisotope phosphorus-32 are readily available. Because of their high specific activity,  $^{32}\text{P}$ -labeled nucleotides are routinely used to label nucleic acids in cell-free systems.

Labeled compounds in which a radioisotope replaces atoms normally present in the molecule have virtually the same chemical properties as the corresponding unlabeled compounds. Enzymes, for instance, generally cannot distinguish between substrates labeled in this way and their unlabeled substrates. The presence of such radioactive atoms is indicated with the isotope in brackets (no hyphen) as a prefix (e.g., [ $^3\text{H}$ ]leucine). In contrast, labeling of almost any biomolecule (e.g., protein or nucleic acid) with the radioisotope iodine-125 ( $^{125}\text{I}$ ) requires the covalent addition of  $^{125}\text{I}$  to a molecule that normally does not have iodine as part of its structure. Because this labeling procedure modifies the chemical structure, the biological activity of the labeled molecule may differ somewhat from that of the unlabeled form. The presence of such radioactive atoms is indicated with the isotope as a prefix followed a hyphen (no bracket) (e.g.,  $^{125}\text{I}$ -trypsin). Standard methods for labeling proteins with  $^{125}\text{I}$  result in covalent attachment of the  $^{125}\text{I}$  primarily to the aromatic rings of tyrosine side chains (mono- and diiodotyrosine). Nonradioactive isotopes are finding increasing use in cell biology, especially in nuclear magnetic resonance studies and in mass spectroscopy applications, as will be explained below.

**Labeling Experiments and Detection of Radiolabeled Molecules** Whether labeled compounds are detected by **autoradiography**—exposure of the sample on a two-dimensional detector (photographic emulsion or electronic detector)—or their radioactivity is measured in an appropriate “counter,” the amount of a radiolabeled compound in a sample can be determined with great precision.

In one use of autoradiography, a tissue, cell, or cell constituent is labeled with a radioactive molecule, unassociated radioactive material is washed away, and the structure of the sample is stabilized either by chemically cross-linking the macromolecules in the sample (“fixation”) or by freezing it. The sample is then overlaid with a photographic emulsion that is sensitive to radiation. Development of the emulsion yields small silver grains whose distribution corresponds to that of the radioactive material and is usually detected by microscopy. Autoradiographic studies of whole

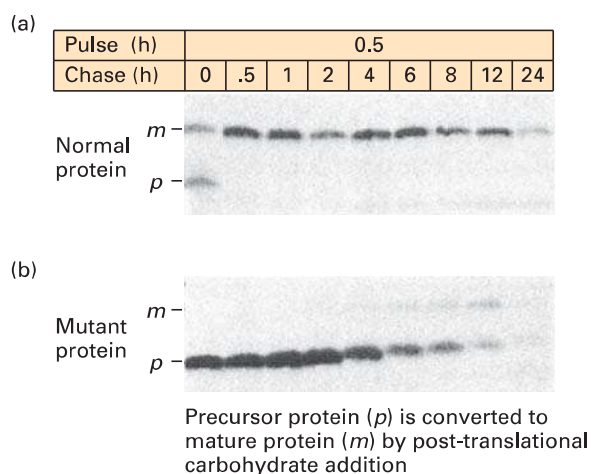
**TABLE 3-1** Radioisotopes Commonly Used in Biological Research

Isotope	Half-Life
Phosphorus-32	14.3 days
Iodine-125	60.4 days
Sulfur-35	87.5 days
Tritium (hydrogen-3)	12.4 years
Carbon-14	5730.4 years

cells were crucial in determining the intracellular sites where various macromolecules are synthesized and the subsequent movements of those macromolecules within cells. Various techniques employing fluorescence microscopy, which we describe in Chapter 4, have largely supplanted autoradiography for studies of this type. However, autoradiography is sometimes used in various assays for detecting specific isolated DNA or RNA sequences at specific tissue locations (see Chapter 6) in a technique referred to as *in situ* hybridization.

Quantitative measurements of the amount of radioactivity in a labeled material are performed with several different instruments. A Geiger counter measures ions produced in a gas by the  $\beta$  particles or  $\gamma$  rays emitted from a radioisotope. These instruments are mostly handheld devices used to monitor radioactivity in the laboratory to protect investigators from excess exposure. In a scintillation counter, a radiolabeled sample is mixed with a liquid containing a fluorescent compound that emits a flash of light when it absorbs the energy of the  $\beta$  particles or  $\gamma$  rays released in the decay of the radioisotope; a phototube in the instrument detects and counts these light flashes. Phosphorimagers detect radioactivity using a two-dimensional array detector, storing digital data on the number of disintegrations per minute per small pixel of surface area. These instruments, which can be thought of as a kind of reusable electronic film, are commonly used to quantify radioactive molecules separated by gel electrophoresis and are replacing photographic emulsions for this purpose.

Combinations of labeling and biochemical techniques and of visual and quantitative detection methods are often employed in labeling experiments. For instance, to identify the major proteins synthesized by a particular cell type, a sample of the cells is incubated with a radiolabeled amino acid (e.g., [ $^{35}\text{S}$ ]methionine) for a few minutes, during which time the labeled amino acid enters the cells and mixes with the cellular pool of unlabeled amino acids, and some of it is biosynthetically incorporated into newly synthesized proteins. Subsequently, unincorporated radiolabeled amino acid is washed away from the cells. The cells are harvested, and the mixture of cellular proteins is extracted from the cells (for example, by a detergent solution) and then separated by any of the methods commonly used to resolve complex protein mixtures into individual components. Gel electrophoresis in combination with autoradiography or phosphorimager analysis is often the method of choice. The radioactive bands in the gel correspond to newly synthesized proteins, which have incorporated the radiolabeled amino acid. To detect a specific protein of interest, rather than the entire ensemble of biosynthetically radiolabeled proteins, a specific protein can be isolated by immunoprecipitation. The precipitate is then solubilized, for example, in an SDS-containing buffer, and the sample is analyzed by SDS-PAGE followed by autoradiography to detect the protein that is radioactively labeled. In this type of experiment, a fluorescent compound that is activated by the radiation (“scintillator”) may be infused into the gel on completion of the electrophoretic separation so that the light emitted can be used to detect the presence of



**EXPERIMENTAL FIGURE 3-42 Pulse-chase experiments can track the pathway of protein modification within cells.** (a) To follow the fate of a specific newly synthesized protein in cells, cells were incubated with [ $^{35}\text{S}$ ]methionine for 0.5 hours (the pulse) to label all newly synthesized proteins, and any radioactive amino acid not incorporated into the cells was then washed away. The cells were further incubated (the chase) for varying times up to 24 hours, and samples from each time of chase were subjected to immunoprecipitation to isolate one specific protein (here the low-density lipoprotein receptor). SDS-PAGE of the immunoprecipitates followed by autoradiography permitted visualization of the target protein, which is initially synthesized as a small precursor (p) and then rapidly modified to a larger mature form (m) by addition of carbohydrates. About half of the labeled protein was converted from p to m during the pulse; the rest was converted after 0.5 hours of chase. The protein remained stable for 6–8 hours before it began to be degraded (as indicated by reduced band intensity). (b) The same experiment was performed in cells in which a mutant form of the protein is made. The mutant p form cannot be properly converted to the m form, and it is more quickly degraded than the normal protein.

[© Kozarsky et al., *The Journal of Cell Biology*, **102**: 1567–1575. doi:10.1083/jcb.102.5.1567]

the labeled protein, using either film or a two-dimensional electronic detector. An example is shown in the experiment described below (Figure 3-42). This method is particularly useful for weak  $\beta$  emitters such as  $^3\text{H}$ .

**Pulse-chase** experiments are particularly useful for tracing changes in the intracellular location of proteins or the modification of a protein or metabolite over time. In this experimental protocol, a cell sample is exposed to a radiolabeled compound that can be incorporated into or otherwise attached to a cellular molecule of interest—the “pulse”—for a brief period. The pulse ends when the unincorporated radiolabeled molecules are washed away and the cells are exposed to a vast excess of the identical, but unlabeled, compound to dilute the radioactivity of any remaining, but unincorporated, radiolabeled compound. This procedure prevents any incorporation of significant amounts of radiolabel after the “pulse” period and initiates the “chase” period (see Figure 3-42). Samples taken periodically during the chase period are assayed to determine the location or chemical form of the radiolabel as a function of time. Pulse-chase



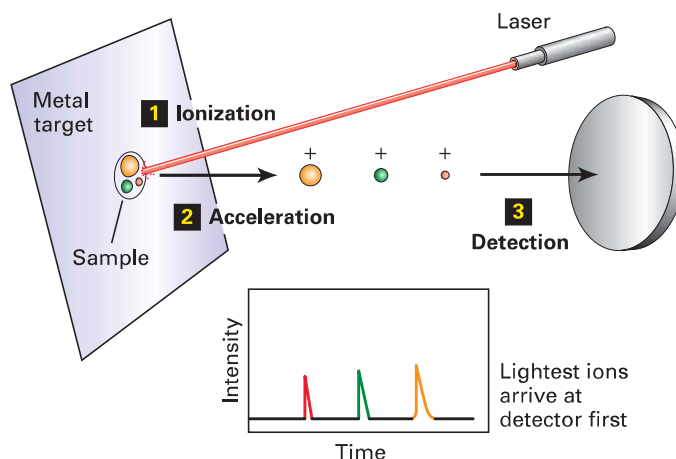
experiments in which the radiolabeled protein is detected by autoradiography after immunoprecipitation and SDS-PAGE are often used to follow the rate of synthesis, modification, and degradation of proteins. In these experiments, radiolabeled amino acid precursors are added during the pulse, and the amounts and characteristics of the radiolabeled target protein are detected during the chase. One can thus observe postsynthetic modifications of the protein, such as the covalent addition of sugars (see Chapters 13 and 14) or proteolytic cleavage, that change its electrophoretic mobility, as well as the rate of degradation of the protein, which is detected as the loss of signal with increasing time of chase. A classic use of the pulse-chase technique with autoradiography was in studies that elucidated the pathway traversed by secreted proteins from their site of synthesis in the endoplasmic reticulum to the cell surface (see Chapter 14).

## Mass Spectrometry Can Determine the Mass and Sequence of Proteins

Mass spectrometry (MS) is a powerful technique for characterizing proteins, especially for determining the mass of a protein or fragments of a protein. With such information in hand, it is also possible to determine part or all of the protein's sequence. This method permits the accurate direct determination of the ratio of the mass ( $m$ ) of a charged molecule (molecular ion) to its charge ( $z$ ), or  $m/z$ . Additional techniques are then used to deduce the absolute mass of the molecular ion.

All mass spectrometers have four key features. The first is an ion source, from which charge, usually in the form of protons, is transferred to the peptide or protein molecules under study (ionization). Their conversion to ions occurs in the presence of a high electric field, which then directs the charged molecular ions into the second key component, the mass analyzer. The mass analyzer, which is always in a high vacuum chamber, physically separates the ions on the basis of their differing mass-to-charge ( $m/z$ ) ratios. The separated ions are subsequently directed to strike a detector, the third key component, which provides a measure of the relative abundances of each of the ions in the sample. The fourth essential component is a computerized data system that is used to calibrate the instrument; to acquire, store, and process the resulting data; and often to direct the instrument to automatically collect additional specific types of data from the sample, based on the initial observations. This type of automated feedback is used for the tandem MS (MS/MS) peptide-sequencing methods described below.

The two most frequently used methods of generating ions of proteins and protein fragments are (1) matrix-assisted laser desorption/ionization (MALDI) and (2) electrospray (ES). In MALDI (Figure 3-43), the peptide or protein sample is mixed with a low-molecular-weight, UV-absorbing organic acid (the matrix) and then dried on a metal target. Energy from a laser ionizes and vaporizes the sample, producing singly charged molecular ions from the constituent molecules. In ES (Figure 3-44a), a sample of peptides or

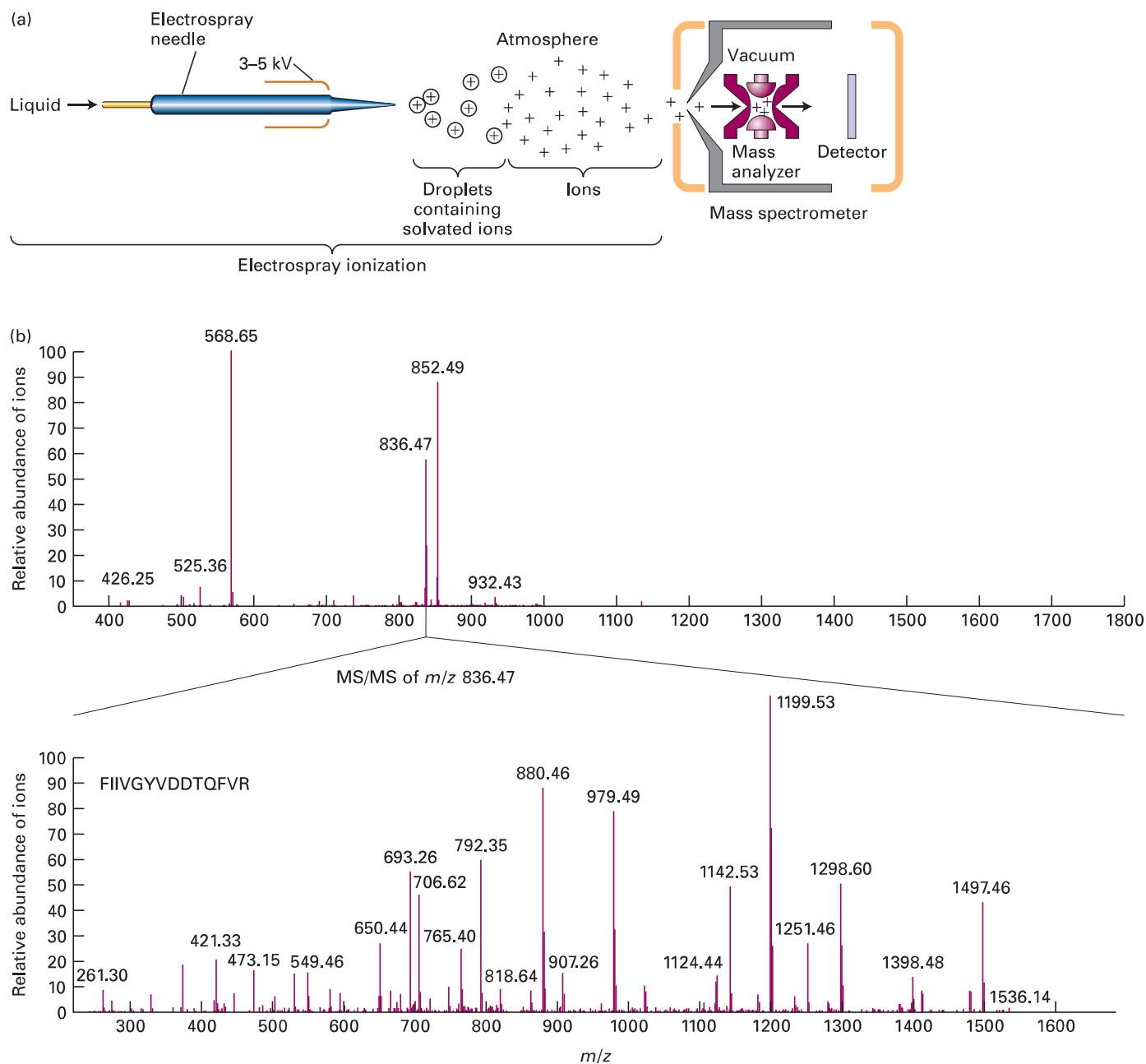


**EXPERIMENTAL FIGURE 3-43 Molecular mass can be determined by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry.** In a MALDI-TOF mass spectrometer, pulses of light from a laser ionize a protein or peptide mixture that is absorbed on a metal target (step 1). An electric field in the mass analyzer accelerates the ions in the sample toward the detector (steps 2 and 3). The time it takes an ion to reach the detector is proportional to the square root of the mass-to-charge ( $m/z$ ) ratio. Among ions having the same charge, the smaller ions move faster (shorter time to the detector). The molecular weight of each ion from the sample is calculated using the time of flight of a standard.

proteins in solution is converted into a fine mist of tiny droplets by spraying through a narrow capillary at atmospheric pressure. The droplets are formed in the presence of a high electric field, which renders them highly charged. The solvent evaporates from the droplets in their short flight (mm) to the entrance of the mass spectrometer's mass analyzer, forming multiply charged ions from the peptides and proteins. The gaseous ions are transferred into the mass analyzer region of the MS, where they are then accelerated by electric fields and separated by the mass analyzer on the basis of their  $m/z$ .

The two most frequently used types of mass analyzers are time-of-flight (TOF) instruments and ion traps. TOF instruments exploit the fact that the time it takes an ion to pass through the length of the mass analyzer before reaching the detector is proportional to the square root of  $m/z$  (smaller ions move faster than larger ones with the same charge; see Figure 3-43). In ion-trap analyzers, tunable electric fields are used to capture, or "trap," ions with a specific  $m/z$  and to sequentially pass the trapped ions out of the mass analyzer onto the detector (see Figure 3-44a). By varying the electric fields, researchers can examine ions with a wide range of  $m/z$  values one by one, producing a mass spectrum, which is a graph of  $m/z$  ( $x$  axis) versus relative abundance, determined by the intensity of the signal measured by the detector ( $y$  axis) (Figure 3-44b, *top panel*).

In tandem, or MS/MS, instruments, any given parent ion in the original mass spectrum (see Figure 3-44b, *top panel*) can be chosen (mass-selected) for further analysis. The chosen ions are transferred into a second chamber in which



**EXPERIMENTAL FIGURE 3-44 Molecular mass of proteins and peptides can be determined by electrospray ionization ion-trap mass spectrometry.** (a) Electrospray (ES) ionization converts proteins and peptides in a solution into highly charged gaseous ions by passing the solution through a needle (forming the droplets) that has a high voltage across it (charging the droplets). Evaporation of the solvent produces gaseous ions that enter a mass spectrometer. The ions are analyzed by an ion-trap mass analyzer that then directs ions to the detector. (b) *Top panel:* Mass spectrum of a mixture of three major and several minor peptides from the mouse H-2 class I histocompatibility antigen Q10  $\alpha$  chain is presented as the relative abundance of the ions striking the detector (y axis) as a function of the mass-to-charge ( $m/z$ ) ratio (x axis). *Bottom panel:* In an MS/MS instrument such as the ion trap

shown in part (a), a specific peptide ion can be selected for fragmentation into smaller ions that are then analyzed and detected. The MS/MS spectrum (also called the product-ion spectrum) provides detailed structural information about the parent ion, including sequence information for peptides. Here the ion with an  $m/z$  of 836.47 was selected and fragmented and the  $m/z$  mass spectrum of the product ions measured. Note there is no longer an ion with an  $m/z$  of 836.47 present because it was fragmented. From the varying sizes of the product ions, the understanding that peptide bonds are often broken in such experiments, the known  $m/z$  values for individual amino acid fragments, and database information, the sequence of the peptide, FIIVGYVDDTQFVR, can be deduced. [Part (b), unpublished data from S. Carr.]

they are broken into smaller fragment ions by collision with an inert gas, and then the  $m/z$  and relative abundances of the resulting fragment ions are measured in a second MS analyzer (Figure 3-44b, *bottom panel*, see also Figure 3-47 later in this chapter). These multiple mass analysis and fragmentation steps all take place within the same machine in about 0.1 seconds per selected parent ion. The fragmentation and subsequent mass analysis permit the sequences of short peptides (<25 amino acids) to be determined because collisional fragmentation occurs primarily at peptide bonds, so the differences in masses between the multiple ion fragments generated correspond to the in-chain masses of the individual amino acids, permitting deduction of the sequence in conjunction with database sequence information (see Figure 3-44b, *bottom panel*).

Mass spectrometry is highly sensitive, able to detect as little as  $1 \times 10^{-16}$  mol (100 attomoles) of a peptide or  $10 \times 10^{-15}$  mol (10 femtomoles) of a protein of 200,000 MW. Errors in mass measurement accuracy are dependent on the specific mass analyzer used, but are typically about 0.01 percent for peptides and 0.05–0.1 percent for proteins. As described in Section 3.6 below, it is possible to use MS to analyze complex mixtures of proteins as well as purified proteins. MS can readily distinguish between two chemically identical peptides that differ only in that one of the peptides contains “heavy” stable (nonradioactive) isotopic forms of one or more elements (e.g., the isotopes  $^2\text{H}$ ,  $^{13}\text{C}$ ,  $^{15}\text{N}$ ) whereas the other contains the most common, “light” isotopes (e.g.,  $^1\text{H}$ ,  $^{12}\text{C}$ ,  $^{14}\text{N}$ ) because the masses of these peptides differ. Most commonly, protein samples are digested by proteases and the peptide digestion products are subjected to analysis. An especially powerful application of MS is to take a complex mixture of proteins from a biological specimen, digest it with trypsin or other proteases, partially separate the components using liquid chromatography, and then transfer the solution flowing out of the chromatographic column directly into an ES tandem mass spectrometer. This technique, called LC-MS/MS, which permits the nearly continuous analysis of a very complex mixture of proteins, will be described in more detail below.

The abundances of ions determined by mass spectrometry in any given sample are relative, not absolute, values. Therefore, if one wants to use MS to compare the amounts of a particular protein in two different samples (e.g., from a normal versus a mutant organism), it is necessary to have an internal standard in the samples—a molecule whose amounts do not differ between the two samples. One then determines the amounts of the protein of interest relative to that of the standard in each sample. This approach permits quantitatively accurate inter-sample comparisons of protein levels. An alternative approach involves simultaneously comparing in a single MS analysis the amounts of proteins from two different cell or tissue samples that are mixed together. This mixing approach is possible provided the proteins in one of the samples contain different stable isotopes than those in the other. The masses of the otherwise chemically identical peptides from the two samples

will then differ (heavy vs. light) and can thus be distinguished by MS. Several methods can be used to chemically or enzymatically incorporate heavy or light isotopes into proteins isolated from cells for such analysis. Alternatively, cells or organisms can first be grown in the presence of amino acids containing either “heavy” or “light” isotope atoms so that these amino acids are biosynthetically incorporated into all the proteins of that sample. Cells are typically incubated with the heavy or light amino acids for five or more cell divisions to ensure that all proteins are thoroughly labeled. Proteins from the two samples are then mixed together, digested into peptides, and the peptides analyzed by mass spectrometry. Proteins and peptides derived from the “heavy” sample can be distinguished in the mass spectrometer from those from the other, “light,” sample because of their higher masses. Thus a direct comparison can be made of the relative amounts of the equivalent proteins in each sample—for example, in tumor versus normal cells or in cells treated with and without a drug. When the samples are cells grown in the laboratory, the method is called *stable isotope labeling with amino acids in cell culture* (SILAC).

### Protein Primary Structure Can Be Determined by Chemical Methods and from Gene Sequences

The classic method for determining the amino acid sequence of a protein is Edman degradation. In this procedure, the free amino group of the N-terminal amino acid of a polypeptide is labeled, and the labeled amino acid is then cleaved from the polypeptide and identified by high-pressure liquid chromatography. The polypeptide is left one residue shorter, with a new amino acid at the N-terminus. The cycle is repeated on the ever-shortening polypeptide until all the residues have been identified.

Before about 1985, biologists commonly used Edman degradation for determining protein sequences. Now, however, complete protein sequences usually are determined primarily by analysis of genome and messenger RNA sequences. The complete genomes of many organisms have already been sequenced, and the database of genome sequences from humans and numerous model organisms is expanding rapidly. As discussed in Chapter 6, the sequences of proteins can be deduced from DNA sequences that are predicted to encode proteins.

A powerful approach for determining the primary structure of an isolated protein combines MS and the use of sequence databases. First, the peptide mass fingerprint of the protein is obtained by MS. A *peptide mass fingerprint* is the list of the molecular weights of peptides that are generated from the protein by digestion with a specific protease, such as trypsin. The molecular weights of the parent protein and its proteolytic fragments are then used to search genome databases for any similar-sized protein with identical or similar peptide mass fingerprints. Mass spectrometry can also be used to directly sequence peptides using MS/MS, as described above.

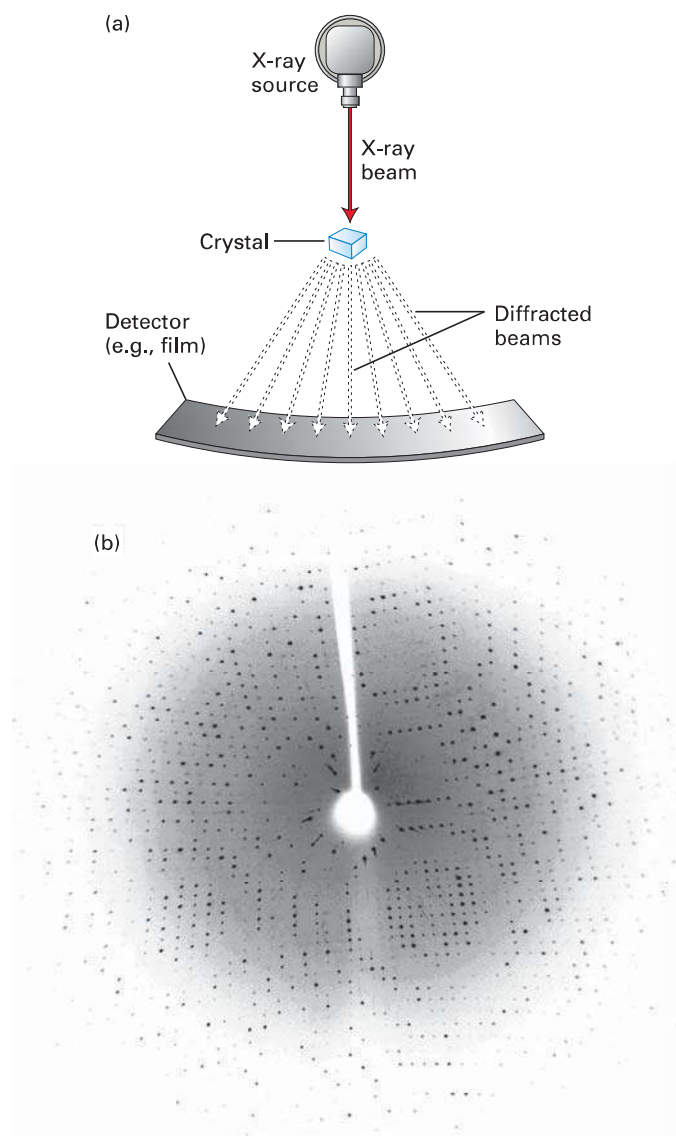


## Protein Conformation Is Determined by Sophisticated Physical Methods

In this chapter, we have emphasized that protein function is dependent on protein structure. Thus, to figure out exactly how a protein works, its three-dimensional structure must be determined. Determining a protein's conformation requires sophisticated physical methods and complex analyses of the experimental data. Here we briefly describe three methods used to generate three-dimensional models of proteins.

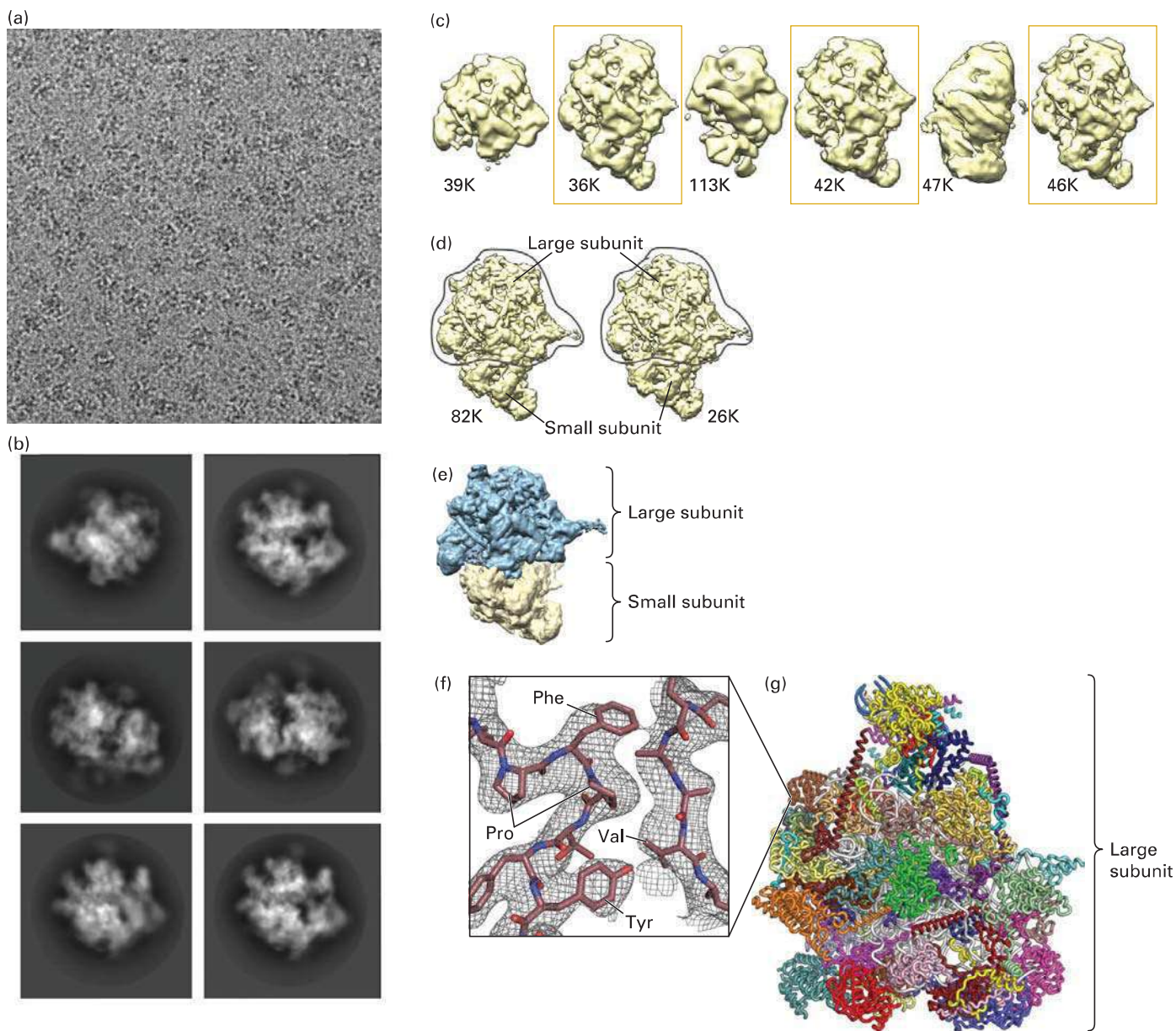
**X-ray Crystallography** The use of x-ray crystallography to determine the three-dimensional structures of proteins was pioneered by Max Perutz and John Kendrew in the 1950s. In this technique, beams of x-rays are passed through a protein crystal, in which millions of protein molecules are precisely aligned with one another in a rigid crystalline array. The wavelengths of x-rays are about 0.1–0.2 nm, short enough to determine the positions of individual atoms in the protein. The electrons in the atoms of the crystal scatter the x-rays, which produce a diffraction pattern of discrete spots when they are intercepted by photographic film or an electronic detector (Figure 3-45). Such patterns are extremely complex—composed of as many as 25,000 diffraction spots, or reflections, whose measured intensities vary depending on the distribution of the electrons in the sample, which is, in turn, determined by the atomic structure and three-dimensional conformation of the protein. Elaborate calculations and modifications of the protein (such as the binding of heavy metals) must be made to interpret the diffraction pattern and calculate the distribution of electrons (called the *electron density map*). A portion of an electron density map of a protein can be seen in Figure 2-9. With the three-dimensional electron density map in hand, one then “fits” a molecular model of the protein to match the electron density, and it is these models that one sees in the various diagrams of proteins throughout this book (e.g., Figure 3-9). The process is analogous to reconstructing the precise shape of a rock from the ripples that it creates when thrown into a pond. Although sometimes the structures of parts of the protein cannot be clearly defined, using x-ray crystallography, researchers are systematically determining the structures of representative types of most proteins. To date, more than 90,000 detailed three-dimensional structures, including more than 35,000 distinct protein sequences, have been established using x-ray crystallography. These structures can be found in the Research Collaboratory for Structural Bioinformatics Protein Data Bank (<http://www.rcsb.org/pdb/home/home.do>), each with its own “PDB” entry.

**Cryoelectron Microscopy** Although some proteins readily crystallize, obtaining crystals of others—particularly large multisubunit proteins and membrane-associated proteins—requires a time-consuming, often robot-assisted trial-and-error effort to find just the right conditions, if they can be found at all. (Growing crystals suitable for structural studies is as much an art as a science.) There are several ways to



**EXPERIMENTAL FIGURE 3-45** X-ray crystallography provides diffraction data from which the three-dimensional structure of a protein can be determined. (a) Basic components of an x-ray crystallographic determination. When a narrow beam of x-rays strikes a crystal, part of it passes straight through and the rest is scattered (diffracted) in various directions. The intensity of the diffracted waves, which form periodic arrangements of diffraction spots, is recorded on an x-ray film or with a solid-state electronic detector. (b) X-ray diffraction pattern for a protein crystal collected on a solid-state detector. From complex analyses of patterns of spots like this one, the locations of the atoms in a protein can be determined. See J. M. Berg, J. L. Tymoczko, G. J. Gatto, and L. Stryer, 2015, *Biochemistry*, 8th ed., Macmillan. [Part (b) courtesy James M. Berger]

determine the structures of such difficult-to-crystallize proteins. One is cryoelectron microscopy (Figure 3-46). In this technique, a dilute protein sample in an aqueous solution is applied in a thin layer to an electron microscope sample holder (a “grid”) and rapidly frozen in liquid helium to preserve its structure. It is then examined in the frozen, hydrated state in a cryoelectron microscope. Images of the protein are



### EXPERIMENTAL FIGURE 3-46 Cryoelectron microscopy analysis of the structure of the human mitochondrial ribosome.

The mitochondrion is a complex, multifunctional intracellular organelle best known for its ability to synthesize the energy carrier ATP (see Chapter 12). Human mitochondria can synthesize proteins encoded by mitochondrial DNA using large (1.7 MDa), multi-protein (at least 78) and multi-RNA complexes called mitochondrial ribosomes that differ somewhat from cytoplasmic ribosomes. (a) Cryoelectron micrograph of isolated human mitochondrial ribosomes. The low contrast between the ribosomes and the buffer solution makes it difficult to clearly see individual, frozen ribosome particles, which are oriented randomly in the image. (b) Automated image processing of 323,292 individual particles permits their grouping into classes based on orientation and averaging of the images within each class to generate clearer images of the ribosome. (c) Additional computational analysis generates distinct structures, each based on tens of thousands of individual particles (the number of particles analyzed for each structure in thousands [K] is shown beneath each). The structures enclosed in boxes were selected for

additional analysis, which produced the two very similar models shown in (d) containing virtually identical large subunits. (e) Color-coded, low-resolution model of the electron density of the large (blue) and small (yellow) subunits. The conformational heterogeneity of the small subunit prevented its high-resolution structure determination from the data shown here. (f) High-magnification view of the experimentally determined electron density (meshwork) from a portion of one of the proteins in the large subunit illustrates how the electron density is used to build the superimposed molecular model of polypeptide chains. In this very small portion of one protein within the large subunit, the side chains of proline (Pro), phenylalanine (Phe), valine (Val), and tyrosine (Tyr) residues are easily seen and demonstrate the power of cryoelectron microscopy to determine protein structures at very high resolutions. (g) Model of the 48 protein subunits (different colors) in the large subunit determined at 3.4 Å resolution. [Republished with permission of American Association for the Advancement of Science, from Brown, A., et al, "Structure of the large ribosomal subunit from human mitochondria." *Science*, 2014, **346** (6210): 718-722; permission conveyed through Copyright Clearance Center, Inc.]



recorded on a very sensitive camera using a low dose of electrons to prevent radiation-induced damage to the structure. Since the individual proteins are in different orientations in the frozen sample, sophisticated computer algorithms analyze the images to sort them into groups with the same orientation. The average image of each orientation is calculated from images of the thousands of different molecules in each group, and then the computer assembles the average images, each of which show views of the protein from different orientations, to reconstruct the protein's structure in three dimensions. Recent advances in this technology have produced structures in which the polypeptide backbone and amino acid side chains can be discerned. These structures help provide insight into the mechanisms underlying the protein's function. The use of cryoelectron microscopy and other types of electron microscopy for visualizing cell structures is discussed in Chapter 4.

**NMR Spectroscopy** The three-dimensional structures of small proteins containing as many as 200 amino acids can be studied routinely with nuclear magnetic resonance (NMR) spectroscopy, and specialized approaches can be used to extend the size range to somewhat larger proteins. In this technique, a concentrated protein solution is placed in a magnetic field, and the effects of different radio frequencies on the nuclear spin states of different atoms are measured. The spin state of any atom is influenced by neighboring atoms in adjacent residues, with closely spaced residues having a greater effect than distant residues. From the magnitude of the effect, the distances between residues can be calculated by a triangulation-like process; these distances are then used to generate a model of the three-dimensional structure of the protein. An important distinction between x-ray crystallography and NMR spectroscopy is that the former method directly determines the locations of the atoms, while the latter directly determines the distances between the atoms, from which the structure is deduced.

Although NMR does not require the crystallization of a protein—a definite advantage—this technique is usually limited to proteins smaller than about 50 kDa (although new techniques permit analysis of the dynamics in much larger proteins). However, NMR analysis can provide information about the ability of a protein to adopt a set of closely related, but not exactly identical, conformations and to move between those conformations (protein dynamics). This is a common feature of proteins, which are not absolutely rigid structures, but can “breathe” or exhibit slight variations in the relative positions of their constituent atoms. In some cases, these variations can have functional significance; for example, they may influence how proteins bind to one another. NMR structural analysis has been particularly useful in studying isolated protein domains, which can often be obtained as stable structures and tend to be small enough for this technique. To date, there are more than 10,000 NMR-determined structures available in the Protein Data Bank.

Another powerful approach to studying protein dynamics and protein-protein interactions is hydrogen/deuterium

exchange mass spectrometry (HXMS). When a protein is placed in a deuterated water ( $D_2O$ ) solution, the rate at which deuterium is exchanged for hydrogen in the amides in the peptide bonds depends on the accessibility of an amide to the solvent. Those amides exposed on the protein's surface are highly accessible and exhibit rapid proton/deuterium exchange. Those amides buried in the center of the protein or in a protein-to-protein interface, as well as those participating in hydrogen bonds with other parts of the protein, exhibit slower proton/deuterium exchange rates. A change in protein conformation or binding to other molecules has the potential to alter the rate of hydrogen/deuterium exchange of one or more amides of a protein. MS analysis permits a hypersensitive assay of such conformational changes, allowing the identification of those parts of the protein that directly bind to other molecules or undergo such conformational changes.

## KEY CONCEPTS OF SECTION 3.5

### Purifying, Detecting, and Characterizing Proteins

- Proteins can be separated from other cell components and from one another on the basis of differences in their physical and chemical properties.
- Centrifugation separates proteins on the basis of their rates of sedimentation, which are influenced by their masses and shapes (see Figure 3-37).
- Electrophoresis separates proteins on the basis of their rates of movement in an applied electric field. SDS-polyacrylamide gel electrophoresis (SDS-PAGE) can resolve polypeptide chains differing in molecular weight by 10 percent or less (see Figure 3-38). Two-dimensional gel electrophoresis provides additional resolution by separating proteins first by charge (first dimension) and then by mass (second dimension).
- Liquid chromatography separates proteins on the basis of their rates of movement through a column packed with spherical beads. Proteins differing in mass are resolved on gel filtration columns; those differing in charge, on ion-exchange columns; and those differing in ligand-binding properties, on affinity columns (see Figure 3-40).
- Various assays are used to detect and quantify proteins. Some assays use a light-producing reaction to generate a readily detected signal. Other assays produce an amplified colored signal with enzymes and chromogenic substrates.
- Antibodies are powerful reagents used to detect, quantify, and isolate proteins.
- Immunoblotting, also called Western blotting, is a frequently used method to study specific proteins that exploits the high specificity and sensitivity of protein detection by



antibodies and the high-resolution separation of proteins by SDS-PAGE (see Figure 3-41).

- Immunoprecipitation, often abbreviated as IP, permits the separation of a protein of interest from other proteins in a complex mixture using antibodies specific for the protein of interest. The antibodies are used to precipitate their target protein out of solution for subsequent analysis. Molecules tightly bound to the target protein can precipitate with it (co-immunoprecipitation).
- Isotopes, both radioactive and nonradioactive, play a key role in the study of proteins and other biomolecules. They can be incorporated into molecules without changing the chemical composition of the molecule or as add-on tags. They can be used to help detect the synthesis, location, processing, and stability of proteins.
- Autoradiography is a technique for detecting radioactively labeled molecules in cells, tissues, or electrophoretic gels using two-dimensional detectors (photographic emulsion or electronic detectors).
- Pulse-chase experiments can determine the intracellular fate of proteins and other metabolites (see Figure 3-42).
- Mass spectrometry is a very sensitive and highly precise method of detecting, identifying, and characterizing proteins and peptides.
- Three-dimensional structures of proteins are obtained by x-ray crystallography, cryoelectron microscopy, and NMR spectroscopy. X-ray crystallography provides the most detailed structures but requires protein crystallization. Cryoelectron microscopy is most useful for large protein complexes, which are difficult to crystallize. Only relatively small proteins are amenable to NMR three-dimensional structural analysis.

## 3.6 Proteomics

For most of the twentieth century, the study of proteins was restricted primarily to the analysis of individual proteins. For example, one would study an enzyme by determining its enzymatic activity (its substrates, products, rate of reaction, requirement for cofactors, pH, etc.), its structure, and its mechanism of action. In some cases, the relationships between a few enzymes that participate in a metabolic pathway might also be studied. On a broader scale, the localization and activity of an enzyme would be examined in the context of a cell or tissue. The effects of mutations, diseases, or drugs on the expression and activity of the enzyme might also be the subject of investigation. This multipronged approach provided deep insight into the function and mechanisms of action of individual proteins or relatively small numbers of interacting proteins. However, such a one-by-one approach to studying proteins does not readily provide a global picture of what is happening in the proteome of a cell, tissue, or entire organism.

## Proteomics Is the Study of All or a Large Subset of Proteins in a Biological System

The advent of genomics (sequencing of genomic DNA and its associated technologies, such as simultaneous analysis of the levels of all mRNAs in cells and tissues) clearly showed that a global, or systems, approach to biology could provide unique and highly valuable insights. Many scientists recognized that a global analysis of the proteins in biological systems had the potential for equally valuable contributions to our understanding. Thus a new field was born—**proteomics**. Proteomics is the systematic study of the amounts, modifications, interactions, localization, and functions of all or subsets of proteins at the whole-organism, tissue, cellular, and subcellular levels.

A number of broad questions are addressed in proteomic studies:

- In a given sample (whole organism, tissue, cell, subcellular compartment), what fraction of the whole proteome is expressed (i.e., which proteins are present)?
- Of those proteins present in the sample, what are their relative abundances?
- What are the relative amounts of the different splice forms and chemically modified forms (e.g., phosphorylated, methylated, fatty acylated) of the proteins?
- Which proteins are present in large multiprotein complexes, and which proteins are in each complex? What are the functions of these complexes, and how do they interact?
- When the state (e.g., growth rate, stage of cell cycle, differentiation, stress level) of a cell changes, do the proteins in the cell, or those secreted from the cell, change in a characteristic (*fingerprint*-like) pattern? Which proteins change, and how (relative amounts, modifications, splice forms, etc.)? [Answering these questions requires a form of *protein expression profiling* that complements the *transcriptional (mRNA) profiling* discussed in Chapter 9.]
- Can such fingerprint-like changes be used for diagnostic purposes? For example, do certain cancers or heart disease cause characteristic changes in blood proteins? Can the proteomic fingerprint help determine if a given cancer is resistant or sensitive to a particular chemotherapeutic drug? [Proteomic fingerprints can also be the starting point for studies of the mechanisms underlying the change of state. Proteins (and other biomolecules) that show changes that are diagnostic of a particular state are called *biomarkers*.]
- Can changes in the proteome help define targets for drugs or suggest mechanisms by which a drug might induce toxic side effects? (If so, it might be possible to engineer modified versions of the drug with fewer side effects.)

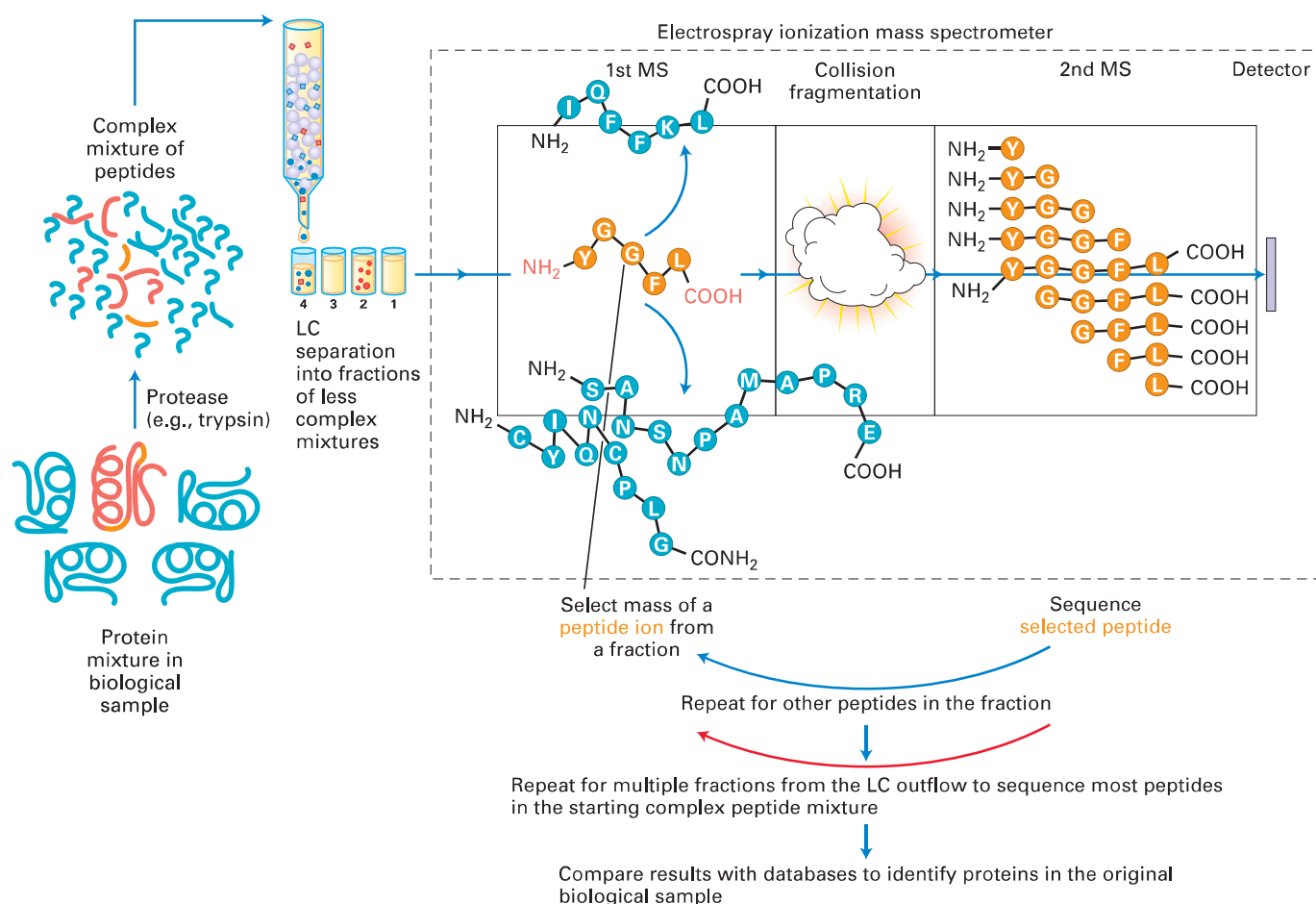
These are just a few of the questions that can be addressed using proteomics. The methods used to answer these questions are as diverse as the questions themselves, and their numbers are growing rapidly.

## Advanced Techniques in Mass Spectrometry Are Critical to Proteomic Analysis

Advances in proteomics technologies (e.g., mass spectrometry) profoundly affect the types of questions that can be practically studied. For many years, two-dimensional gel electrophoresis allowed researchers to separate, display, and characterize complex mixtures of proteins (see Figure 3-39). The spots on a two-dimensional electrophoresis gel could be excised, the protein fragmented by proteolysis (e.g., by trypsin digestion), and the fragments identified by MS. An alternative to this two-dimensional gel electrophoresis method is *high-throughput* LC-MS/MS. Figure 3-47 outlines the general LC-MS/MS approach, in which a complex mixture of proteins is digested with a protease; the myriad resulting peptides are fractionated by LC into multiple, less complex fractions, which are slowly but continuously injected by electrospray ionization into a tandem mass spectrometer. The fractions are then sequentially subjected to multiple

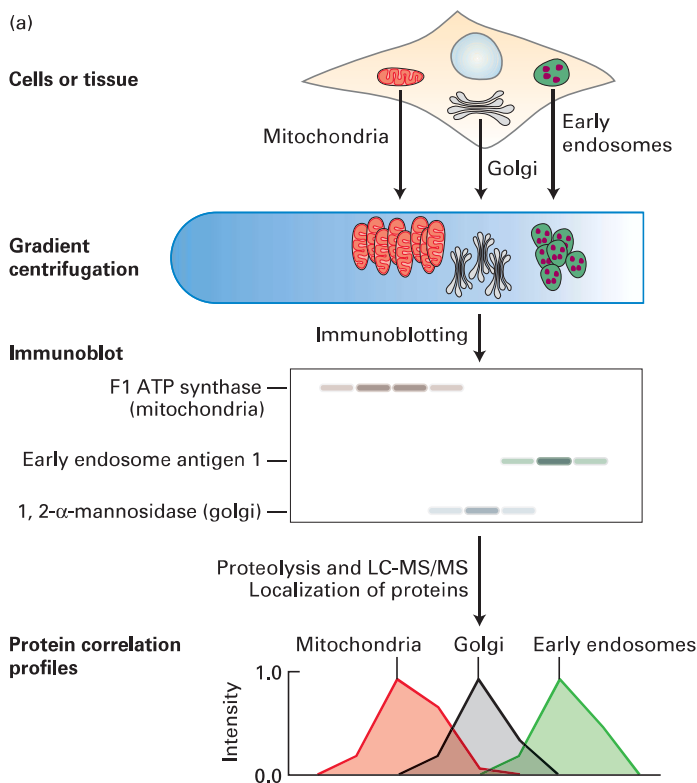
cycles of MS/MS until the sequences of many of the peptides have been determined and used to identify the proteins in the original biological sample. Detection of a substantial fraction of the proteins in whole cells or tissues currently requires samples containing more than 50  $\mu\text{g}$  of protein, an amount equivalent to the protein content of some 70,000–200,000 mammalian cells. Efforts are under way to increase the sensitivity of the method so that eventually one might be able to analyze the proteome of an individual cell.

An example of the use of LC-MS/MS to identify many of the proteins in each organelle is seen in Figure 3-48. Cells from murine (mouse) liver tissue were mechanically broken to release the organelles, and the organelles were partially separated by density-gradient centrifugation. The locations of the organelles in the gradient were determined using immunoblotting with antibodies that recognized previously identified, organelle-specific proteins. Fractions from the gradient were then subjected to LC-MS/MS to identify the proteins in each fraction, and the distributions in the gradient of many



**EXPERIMENTAL FIGURE 3-47 LC-MS/MS is used to identify the proteins in a complex biological sample.** A complex mixture of proteins in a biological sample (e.g., an isolated preparation of Golgi organelles) is digested with a protease. The mixture of resulting peptides is fractionated by liquid chromatography (LC) into multiple, less complex, fractions, which are slowly but continuously injected by

electrospray ionization into a tandem mass spectrometer. The fractions are then sequentially subjected to multiple cycles of MS/MS until the masses and sequences of many of the peptides have been determined and used to identify the proteins in the original biological sample through comparison with protein databases.

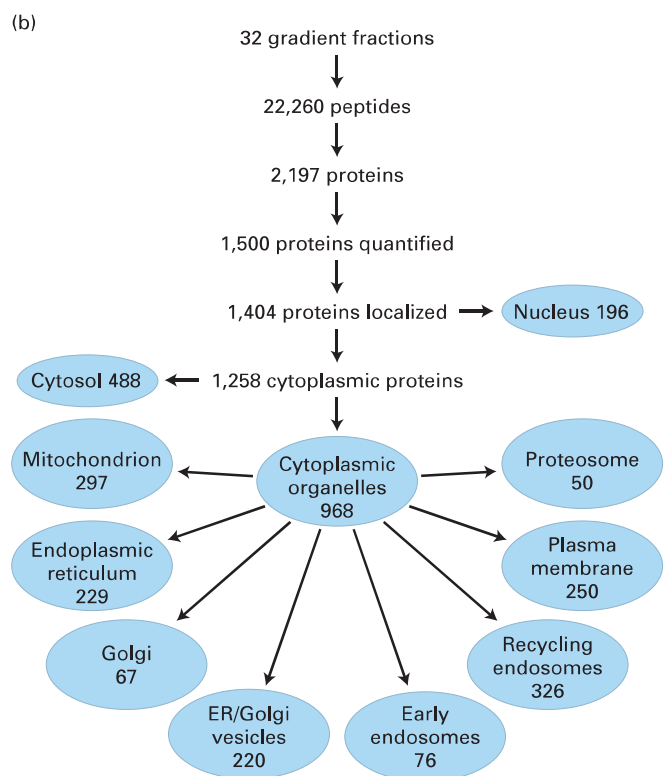


**EXPERIMENTAL FIGURE 3-48 Density-gradient centrifugation and LC-MS/MS can be used to identify many of the proteins in organelles.** (a) The cells in liver tissue were mechanically broken to release the organelles, and the organelles were partially separated by density-gradient centrifugation. The locations of the organelles—which were spread out through the gradient and somewhat overlapped with one another—were determined using immunoblotting with antibodies that recognized previously identified, organelle-specific proteins. Fractions from the gradient were

individual proteins were compared with the distributions of the organelles. This strategy permitted the assignment of many individual proteins to one or more organelles (organelle proteome profiling). More recently, a combination of organelle purification, MS, biochemical localization, and computational methods has been used to show that at least a thousand distinct proteins are localized in the mitochondria of humans and mice.

Proteomics methods combined with molecular genetics methods are currently being used to identify all the protein complexes in eukaryotic cells. For example, in the yeast *Saccharomyces cerevisiae*, approximately 500 complexes, with an average of 4.9 distinct proteins per complex, have been identified. These complexes, in turn, are involved in at least 400 complex-to-complex interactions. Such systematic proteomic studies are providing new insights into the organization of proteins within cells and into how proteins work together to permit cells to live and function.

Phosphoproteomics, the identification and quantification of phosphorylation sites on the proteins in a complex



subjected to proteolysis and LC-MS/MS to identify the peptides, and hence the proteins, in each fraction. Comparisons with the locations of the organelles in the gradient (called protein correlation profiling) permitted assignment of many individual proteins to one or more organelles (organelle proteome identification). (b) The hierarchical breakdown of data derived from the procedures in part (a). Note that not all proteins identified could be assigned to organelles and that some proteins were assigned to more than one organelle. [Data from L. J. Foster et al., 2006, *Cell* **125**:187–199.]

mixture, is playing a growing role in the analysis of cell metabolism and regulation. As we have already learned, the reversible phosphorylation of proteins by kinases and phosphatases is a key mechanism for regulating proteins in cells. Phosphoproteomics permits the simultaneous determination of the phosphorylation states of many proteins and thus provides an important tool for analyzing complex cellular regulatory networks. Only a fraction—in some cases, only a small fraction—of a particular protein might be phosphorylated. Thus phosphoproteomic analysis can require 50–100 times more initial cell or tissue sample material (from about 2.5 to more than 20 mg of total cellular protein per sample) than does standard proteomic analysis. As a consequence, investigators usually use affinity chromatography methods with either metal-containing (e.g.,  $\text{Fe}^{3+}$  or  $\text{TiO}_2$ ) or antibody-containing columns to separate phosphopeptides from nonphosphorylated peptides (phosphopeptide enrichment) prior to subjecting the phosphopeptides to LC-MS/MS analysis.



## KEY CONCEPTS OF SECTION 3.6

### Proteomics

- Proteomics is the systematic study of the amounts (and changes in the amounts), modifications, interactions, localization, and functions of all or subsets of all proteins in biological systems at the whole-organism, tissue, cellular, and subcellular levels.
- Proteomics provides insights into the fundamental organization of proteins within cells and how that organization is influenced by the state of the cells (e.g., differentiation into distinct cell types; responses to stress, disease, and drugs).
- A wide variety of methods are used for proteomic analyses, including two-dimensional gel electrophoresis, density-gradient centrifugation, and mass spectrometry (particularly LC-MS/MS).
- Proteomics has helped begin to identify the proteomes of organelles (“organelle proteome profiling”) as well as the organization of individual proteins into multiprotein complexes (see Figure 3-48).
- Phosphoproteomics is a specialized application of proteomics that identifies the collection of phosphorylated proteins (phosphoproteome) in cells and characterizes how the level of phosphorylation of these proteins varies as the state of the cells changes.



**LaunchPad**  
macmillan learning

Visit LaunchPad to access study tools and to learn more about the content in this chapter.

- Perspectives for the Future
- Analyze the Data
- Extended References
- Additional study tools, including videos, animations, and quizzes

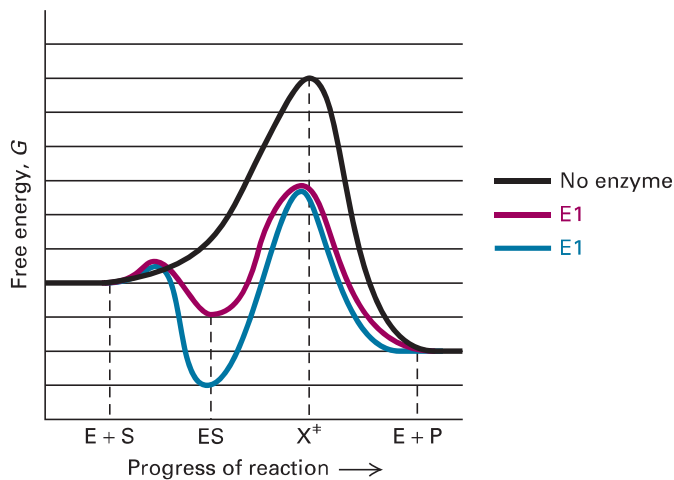
### Key Terms

active site	91	$\beta$ turn	70
allostery	100	chaperone	83
$\alpha$ helix	70	conformation	67
amyloid fibril	87	cooperativity	100
autoradiography	114	domain	76
$\beta$ sheet	70	electrophoresis	107

enzyme	90	proteome	68
homology	80	proteomics	122
kinase	103	quaternary structure	78
$K_m$	92	rate-zonal	
ligand	89	centrifugation	106
liquid		secondary structure	70
chromatography	110	structural motif	75
peptide bond	69	tertiary structure	72
phosphorylation	103	ubiquitin	99
polypeptide	70	$V_{max}$	91
primary structure	70	Western blotting	112
proteasome	97	x-ray	
protein	70	crystallography	119

### Review the Concepts

1. The three-dimensional structure of a protein is determined by its primary, secondary, and tertiary structures. Define the *primary*, *secondary*, and *tertiary structures*. What are some of the common secondary structures? What are the forces that hold together the secondary and tertiary structures?
2. Proper folding of proteins is essential for their biological activity. In general, the functional conformation of a protein is the conformation with lowest energy. This means that if an unfolded protein is allowed to reach equilibrium, it should assemble automatically into its native, functioning folded state. Why then is there a need for molecular chaperones and chaperonins in cells? What different roles do molecular chaperones and chaperonins play in the folding of proteins?
3. Enzymes catalyze chemical reactions. What constitutes the active site of an enzyme? What are the turnover number ( $k_{cat}$ ), the Michaelis constant ( $K_m$ ), and the maximal velocity ( $V_{max}$ ) of an enzyme? The  $k_{cat}$  (catalytic rate constant) for carbonic anhydrase is  $5 \times 10^5$  molecules per second. This is a “rate constant,” but not a “rate.” What is the difference? By what concentration would you multiply this rate constant in order to determine an actual rate of product formation ( $V$ )? Under what circumstances would this rate become equal to the maximal velocity ( $V_{max}$ ) of the enzyme?
4. The following reaction coordinate diagram charts the energy of a substrate molecule (S) as it passes through a transition state ( $X^\ddagger$ ) on its way to becoming a stable product (P) alone or in the presence of one of two different enzymes (E1 and E2). How does the addition of either enzyme affect the change in Gibbs free energy ( $\Delta G$ ) for the reaction? Which of the two enzymes binds with greater affinity to the substrate? Which enzyme better stabilizes the transition state? Which enzyme functions as a better catalyst?



5. A healthy immune system can raise antibodies that recognize and bind with high affinity to almost any stable molecule. The molecule to which an antibody binds is known as an antigen. Antibodies have been exploited by enterprising scientists to generate valuable tools for research, diagnosis, and therapy. One clever application is the generation of antibodies that function like enzymes to catalyze complicated chemical reactions. If you wished to produce such a “catalytic” antibody, what would you suggest using as the antigen? Should it be the substrate of the reaction? The product? Something else?

6. Proteins are degraded in cells. What is ubiquitin, and what role does it play in tagging proteins for degradation? What is the role of proteasomes in protein degradation? How might proteasome inhibitors serve as chemotherapeutic (cancer-treating) agents?

7. The function of proteins can be regulated in a number of ways. What is cooperativity, and how does it influence protein function? Describe how protein phosphorylation and proteolytic cleavage can modulate protein function.

8. A number of techniques can separate proteins on the basis of their differences in mass. Describe the use of two of these techniques, centrifugation and gel electrophoresis. The blood proteins transferrin (MW 76 kDa) and lysozyme (MW 15 kDa) can be separated by rate-zonal centrifugation or SDS-polyacrylamide gel electrophoresis. Which of the two proteins will sediment faster during centrifugation? Which will migrate faster during electrophoresis?

9. Liquid chromatography is an analytical method used to separate proteins. Describe the principles for separating proteins by gel filtration, ion-exchange, and affinity chromatography.

10. Various methods have been developed for detecting proteins. Describe how radioisotopes and autoradiography can be used for labeling and detecting proteins. How does Western blotting detect proteins?

11. Physical methods are often used to determine protein conformation. Describe how x-ray crystallography, cryo-electron microscopy, and NMR spectroscopy can be used to determine the shapes of proteins. What are the advantages

and disadvantages of each method? Which is better for small proteins? Large proteins? Huge macromolecular assemblies?

12. Mass spectrometry is a powerful tool in proteomics. What are the four key features of a mass spectrometer? Describe briefly how MALDI and two-dimensional polyacrylamide gel electrophoresis could be used to identify a protein expressed in cancer cells but not in normal healthy cells.

## References

### Web Sites

- Entry site into proteins, structures, genomes, and taxonomy: <http://www.ncbi.nlm.nih.gov/Entrez/>
- The protein 3-D structure database: <http://www.rcsb.org/>
- Structural classifications of proteins: <http://scop.berkeley.edu/>
- Sites containing general information about proteins: <http://www.expasy.ch/>; <http://www.proweb.org/>; <http://scop.berkeley.edu/intro.html>
- PROSITE database of protein families and domains: <http://www.expasy.org/prosite/>
- Domain organization of proteins and large collection of multiple sequence alignments: <http://www.sanger.ac.uk/Software/Pfam/>; <http://people.cryst.bbk.ac.uk/ubcg16z/cpn/elmovies.html>
- MitoCarta: An Inventory of Mammalian Mitochondrial Genes: <http://www.broadinstitute.org/pubs/MitoCarta/index.html>
- Human protein atlas with expression of proteins in different tissues: <http://www.proteinatlas.org/>

### Hierarchical Structure of Proteins

- Dunker, A. K., et al. 2015. Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.* 37:44–55.
- Levitt, M. 2009. Nature of the protein universe. *P. Natl. Acad. Sci. USA* 106:11079–11084.
- Patthy, L. 1999. *Protein Evolution*. Blackwell Science.
- Vogel, C., and C. Chothia. 2006. Protein family expansions and biological complexity. *PLoS Comput. Biol.* 2(5):e48.
- Yaffe, M. B. 2006. “Bits” and pieces. *Sci. STKE* 2006:pe28.

### Protein Folding

- Brandvold, K. R., and R. I. Morimoto. 2015. The chemical biology of molecular chaperones—implications for modulation of proteostasis. *J. Mol. Biol.* 427:2931–2947.
- Coulson, A. F., and J. Moul. 2002. A unfold, mesofold, and superfold model of protein fold use. *Proteins* 46:61–71.
- Daggett, V., and A. R. Fersht. 2003. Is there a unifying mechanism for protein folding? *Trends Biochem. Sci.* 28:18–25.
- Dobson, C. M. 1999. Protein misfolding, evolution, and disease. *Trends Biochem. Sci.* 24:329–332.
- Jackrel, M. E., et al. 2014. Potentiated Hsp104 variants antagonize diverse proteotoxic misfolding events. *Cell* 156:170–182.
- Knowles, T. P., M. Vendruscolo, and C. M. Dobson. 2014. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* 15:384–396.
- Lavery, L. A., et al. 2014. Structural asymmetry in the closed state of mitochondrial Hsp90 (TRAP1) supports a two-step ATP hydrolysis mechanism. *Mol. Cell* 53:330–343.
- Saibil, H. 2013. Chaperone machines for protein folding, unfolding and disaggregation. *Nat. Rev. Mol. Cell Biol.* 14:630–642.
- Schmidpeter, P. A., and F. X. Schmid. 2015. Prolyl isomerization and its catalysis in protein folding and protein function. *J. Mol. Biol.* 427:1609–1631.

Taipale, M., D. F. Jarosz, and S. Lindquist. 2010. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nat. Rev. Mol. Cell Biol.* **11**:515–528.

Valastyan, J. S., and S. Lindquist. 2014. Mechanisms of protein-folding diseases at a glance. *Dis. Model Mech.* **7**:9–14.

### Protein Binding and Enzyme Catalysis

Fersht, A. 1999. *Enzyme Structure and Mechanism*, 3d ed. W. H. Freeman and Company.

Martínez Cuesta, S., et al. 2015. The classification and evolution of enzyme function. *Biophys. J.* **109**:1082–1086.

Radisky, E. S., et al. 2006. Insights into the serine protease mechanism from atomic resolution structures of trypsin reaction intermediates. *P. Natl. Acad. Sci. USA* **103**:6835–6840.

### Regulating Protein Function

Bellelli, A., et al. 2006. The allosteric properties of hemoglobin: insights from natural and site directed mutants. *Curr. Prot. Pep. Sci.* **7**:17–45.

Campbell, M. G., et al. 2015. 2.8 Å resolution reconstruction of the *Thermoplasma acidophilum* 20S proteasome using cryo-electron microscopy. *eLife*. 10.7554/eLife.06380.

Glickman, M. H., and A. Ciechanover. 2002. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol. Rev.* **82**:373–428.

Goldberg, A. L., S. J. Elledge, and J. W. Harper. 2001. The cellular chamber of doom. *Sci. Am.* **284**:68–73.

Goldberg, A. L. 2003. Protein degradation and protection against misfolded or damaged proteins. *Nature* **426**:895–899.

Kern, D., and E. R. Zuiderweg. 2003. The role of dynamics in allosteric regulation. *Curr. Opin. Struc. Biol.* **13**:748–757.

Kisselev, A. F., A. Callard, and A. L. Goldberg. 2006. Importance of the different proteolytic sites of the proteasome and the efficacy of inhibitors varies with the protein substrate. *J. Biol. Chem.* **281**:8582–8590.

Lim, W. A. 2002. The modular logic of signaling proteins: building allosteric switches from simple binding domains. *Curr. Opin. Struc. Biol.* **12**:61–68.

Sahtoe, D. D., and T. K. Sixma. 2015. Layers of DUB regulation. *Trends Biochem. Sci.* **40**(8):456–467.

Sowa, M. E., et al. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell* **138**:389–403.

### Purifying, Detecting, and Characterizing Proteins

Engen, J. R., et al. 2013. Partial cooperative unfolding in proteins as observed by hydrogen exchange mass spectrometry. *Int. Rev. Phys. Chem.* **32**:96–127.

Hames, B. D. *A Practical Approach*. Oxford University Press. A methods series that describes protein purification methods and assays.

Liao, M., et al. 2014. Single particle electron cryo-microscopy of a mammalian ion channel. *Curr. Opin. Biol.* **27**:1–7.

Nogales, E., and S. H. Scheres. 2015. Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Mol. Cell* **58**:677–689.

Rosenzweig, R., and L. E. Kay. 2014. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu. Rev. Biochem.* **83**:291–315.

Zhang, G., et al. 2014. Overview of peptide and protein analysis by mass spectrometry. *Curr. Protoc. Mol. Biol.* **108**:10.21.1–10.21.30.

### Proteomics

Azimifar, S. B., et al. 2014. Cell-type-resolved quantitative proteomics of murine liver. *Cell Metab.* **20**:1076–1087.

Calvo, S. E., and V. K. Mootha. 2010. The mitochondrial proteome and human disease. *Annu. Rev. Genomics Hum. Genet.* **11**:25–34.

Cox, J., and M. Mann. 2011. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80**:273–299.

Foster, L. J., et al. 2006. A mammalian organelle map by protein correlation profiling. *Cell* **125**:187–199.

Krogan, N. J., et al. 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**:637–643.

Rifai, N., M. A. Gillette, and S. A. Carr. 2006. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnol.* **24**:971–983.

Roux, P. P., and P. Thibault. 2013. The coming of age of phosphoproteomics—from large data sets to inference of protein functions. *Mol Cell Proteomics*:**12**:3453–3464.

Walther, T. C., and M. Mann. 2010. Mass spectrometry-based proteomics in cell biology. *J. Cell Biol.* **190**:491–500.