Chapter 3

Proteins

In This Chapter
The Shape and Structure of Proteins
Protein function

When we look at a cell through a microscope or analyze its electrical or biochemical activity, we are, in essence, observing proteins. Proteins constitute most of a cell’s dry mass. They are not only the cell’s building blocks; they also execute the majority of the cell’s functions. Thus, proteins that are enzymes provide the intricate molecular surfaces inside a cell that catalyze its many chemical reactions. Proteins embedded in the plasma membrane form channels and pumps that control the passage of small molecules into and out of the cell. Other proteins carry messages from one cell to another, or act as signal integrators that relay sets of signals inward from the plasma membrane to the cell nucleus. Yet others serve as tiny molecular machines with moving parts: kinesin, for example, propels organelles through the cytoplasm; topoisomerase can untangle knotted DNA molecules. Other specialized proteins act as antibodies, toxins, hormones, antifreeze molecules, elastic fibers, ropes, or sources of luminescence. Before we can hope to understand how genes work, how muscles contract, how nerves conduct electricity, how embryos develop, or how our bodies function, we must attain a deep understanding of proteins.

THE SHAPE AND STRUCTURE OF PROTEINS

From a chemical point of view, proteins are by far the most structurally complex and functionally sophisticated molecules known. This is perhaps not surprising, once we realize that the structure and chemistry of each protein has been developed and fine-tuned over billions of years of evolutionary history. The theoretical calculations of population geneticists reveal that, over evolutionary time periods, a surprisingly small selective advantage is enough to cause a randomly altered protein sequence to spread through a population of organisms. Yet, even to experts, the remarkable versatility of proteins can seem truly amazing.

In this section, we consider how the location of each amino acid in the long string of amino acids that forms a protein determines its three-dimensional shape. Later in the chapter, we use this understanding of protein structure at the atomic level to describe how the precise shape of each protein molecule determines its function in a cell.

The Shape of a Protein Is Specified by Its Amino Acid Sequence

There are 20 different of amino acids in proteins that are coded for directly in an organism’s DNA, each with different chemical properties. A protein molecule is made from a long unbranched chain of these amino acids, each linked to its neighbor through a covalent peptide bond. Proteins are therefore also known as polypeptides. Each type of protein has a unique sequence of amino acids, and there are many thousands of different proteins in a cell.

The repeating sequence of atoms along the core of the polypeptide chain is referred to as the polypeptide backbone. Attached to this repetitive chain are those portions of the amino acids that are not involved in making a peptide bond and that give each amino acid its unique properties: the 20 different amino acid side chains (Figure 3–1). Some of these side chains are nonpolar and hydrophobic (“water-fearing”), others are negatively or positively charged, some readily form covalent bonds, and so on. Panel 3–1 (pp. 112–113) shows their atomic structures and Figure 3–2 lists their abbreviations.

As discussed in Chapter 2, atoms behave almost as if they were hard spheres with a definite radius (their van der Waals radius). The requirement that no two atoms overlap plus other constraints limit the possible bond angles in a polypeptide chain (Figure 3–3), severely restricting the possible three-dimensional arrangements (or conformations) of atoms. Nevertheless, a long flexible chain such as a protein can still fold in an enormous number of ways.

The folding of a protein chain is also determined by many different sets of weak noncovalent bonds that form between one part of the chain and another. These involve atoms in the polypeptide backbone, as well as atoms in the amino acid side chains. There are three types of these weak bonds: hydrogen bonds, electrostatic attractions, and van der Waals attractions, as explained in Chapter 2 (see p. 44). Individual noncovalent bonds are 30–300 times weaker than the typical covalent bonds that create biological molecules. But many weak bonds acting in parallel can hold two regions of a polypeptide chain tightly together. In this way, the combined strength of large numbers of such noncovalent bonds determines the stability of each folded shape (Figure 3–4).

A fourth weak force—a hydrophobic clustering force—also has a central role in determining the shape of a protein. As described in Chapter 2, hydrophobic molecules, including the nonpolar side chains of particular amino acids, tend to be forced together in an aqueous environment in order to minimize their disruptive effect on the hydrogen-bonded network of water molecules (see Panel 2–2, pp. 92–93). Therefore, an important factor governing the folding of any protein is the distribution of its polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein—belonging to such amino acids as phenylalanine, leucine, valine, and tryptophan—tend to cluster in the interior of the molecule (just as hydrophobic oil droplets coalesce in water to form one large droplet).

This enables them to avoid contact with the water that surrounds them inside a cell. In contrast, polar groups—such as those belonging to arginine, glutamine, and histidine—tend to arrange themselves near the outside of the molecule, where they can form hydrogen bonds with water and with other polar molecules (Figure 3–5). Polar amino acids buried within the protein are usually hydrogen-bonded to other polar amino acids or to the polypeptide backbone.

Proteins Fold into a Conformation of Lowest Energy

As a result of all of these interactions, most proteins have a particular three-dimensional structure, which is determined by the order of the amino acids in its chain. The final folded structure, or conformation, of any polypeptide chain is generally the one that minimizes its free energy. Biologists have studied protein folding in a test tube using highly purified proteins. Treatment with certain solvents, which disrupt the noncovalent interactions holding the folded chain together, unfolds, or denatures, a protein. This treatment converts the protein into a flexible polypeptide chain that has lost its natural shape. When the denaturing solvent is removed, the protein often refolds spontaneously, or renatures, into its original conformation. This indicates that the amino acid sequence contains all of the information needed for specifying the three-dimensional shape of a protein, a critical point for understanding cell biology.

Most proteins fold up into a single stable conformation. However, this conformation changes slightly when the protein interacts with other molecules in the cell. This change in shape is often crucial to the function of the protein, as we see later.

Although a protein chain can fold into its correct conformation without outside help, in a living cell special proteins called molecular chaperones often assist in protein folding. Molecular chaperones bind to partly folded polypeptide chains and help them progress along the most energetically favorable folding pathway. In the crowded conditions of the cytoplasm, chaperones are required to prevent the temporarily exposed hydrophobic regions in newly synthesized protein chains from associating with each other to form protein aggregates (see p. 355).

However, the final three-dimensional shape of the protein is still specified by its amino acid sequence: chaperones simply make reaching the folded state more reliable.

Proteins come in a wide variety of shapes, and most are between 50 and 2000 amino acids long. Large proteins usually consist of several distinct protein domains—structural units that fold more or less independently of each other, as we discuss below. The structure of even a small domain is complex, and for clarity, several different representations are conventionally used, each of which emphasizes distinct features. As an example, Figure 3–6 presents four representations of a protein domain called SH2, a structure present in many different proteins in eukaryotic cells and involved in cell signaling (see Figure 15–46).

Descriptions of protein structures are aided by the fact that proteins are built up from combinations of several common structural motifs, as we discuss next.

The α Helix and the β Sheet Are Common Folding Patterns

When we compare the three-dimensional structures of many different protein molecules, it becomes clear that, although the overall conformation of each protein is unique, two regular folding patterns are often found within them. Both patterns were discovered more than 60 years ago from studies of hair and silk. The first folding pattern to be discovered, called the α helix, was found in the protein α-keratin, which is abundant in skin and its derivatives—such as hair, nails, and horns. Within a year of the discovery of the α helix, a second folded structure, called a β sheet, was found in the protein fibroin, the major constituent of silk. These two patterns are particularly common because they result from hydrogen-bonding between the N–H and C=O groups in the polypeptide backbone, without involving the side chains of the amino acids. Thus, although incompatible with some amino acid side chains, many different amino acid sequences can form them. In each case, the protein chain adopts a regular, repeating conformation. Figure 3–7 illustrates the detailed structures of these two important conformations, which in ribbon models of proteins are represented by a helical ribbon and by a set of aligned arrows, respectively.

The cores of many proteins contain extensive regions of β sheet. As shown in Figure 3–8, these β sheets can form either from neighboring segments of the polypeptide backbone that run in the same orientation (parallel chains) or from a polypeptide backbone that folds back and forth upon itself, with each section of the chain running in the direction opposite to that of its immediate neighbors (antiparallel chains). Both types of β sheet produce a very rigid structure, held together by hydrogen bonds that connect the peptide bonds in neighboring chains (see Figure 3–7C).

An α helix is generated when a single polypeptide chain twists around on itself to form a rigid cylinder. A hydrogen bond forms between every fourth peptide bond, linking the C=O of one peptide bond to the N–H of another (see Figure 3–7A). This gives rise to a regular helix with a complete turn every 3.6 amino acids. The SH2 protein domain illustrated in Figure 3–6 contains two α helices, as well as a three-stranded antiparallel β sheet.

Regions of α helix are abundant in proteins located in cell membranes, such as transport proteins and receptors. As we discuss in Chapter 10, those portions of a transmembrane protein that cross the lipid bilayer usually cross as α helices composed largely of amino acids with nonpolar side chains. The polypeptide backbone, which is hydrophilic, is hydrogen-bonded to itself in the α helix and shielded from the hydrophobic lipid environment of the membrane by its protruding nonpolar side chains (see also Figure 3–75A).

In other proteins, α helices wrap around each other to form a particularly stable structure, known as a coiled-coil. This structure can form when the two (or in some cases, three or four) α helices have most of their nonpolar (hydrophobic) side chains on one side, so that they can twist around each other with these side chains facing inward (Figure 3–9). Long rodlike coiled-coils provide the structural framework for many elongated proteins. Examples are α-keratin, which forms the intracellular fibers that reinforce the outer layer of the skin and its appendages, and the myosin molecules responsible for muscle contraction.

Protein Domains Are Modular Units from Which Larger Proteins Are Built

Even a small protein molecule is built from thousands of atoms linked together by precisely oriented covalent and noncovalent bonds. Biologists are aided in visualizing these extremely complicated structures by various graphic and computer-based three-dimensional displays. The student resource site that accompanies this book contains computer-generated images of selected proteins, displayed and rotated on the screen in a variety of formats.

Scientists distinguish four levels of organization in the structure of a protein. The amino acid sequence is known as the primary structure. Stretches of polypeptide chain that form α helices and β sheets constitute the protein’s secondary structure. The full three-dimensional organization of a polypeptide chain is sometimes referred to as the tertiary structure, and if a particular protein molecule is formed as a complex of more than one polypeptide chain, the complete structure is designated as the quaternary structure.

Studies of the conformation, function, and evolution of proteins have also revealed the central importance of a unit of organization distinct from these four. This is the protein domain, a substructure produced by any contiguous part of a polypeptide chain that can fold independently of the rest of the protein into a compact, stable structure. A domain usually contains between 40 and 350 amino acids, and it is the modular unit from which many larger proteins are constructed.

The different domains of a protein are often associated with different functions. Figure 3–10 shows an example—the Src protein kinase, which functions in signaling pathways inside vertebrate cells (Src is pronounced “sarc”). This protein is considered to have three domains: the SH2 and SH3 domains have regulatory roles, while the C-terminal domain is responsible for the kinase catalytic activity. Later in the chapter, we shall return to this protein, in order to explain how proteins can form molecular switches that transmit information throughout cells.

Figure 3–11 presents ribbon models of three differently organized protein domains. As these examples illustrate, the central core of a domain can be constructed from α helices, from β sheets, or from various combinations of these two fundamental folding elements.

The smallest protein molecules contain only a single domain, whereas larger proteins can contain several dozen domains, often connected to each other by short, relatively unstructured lengths of polypeptide chain that can act as flexible hinges between domains.

Few of the Many Possible Polypeptide Chains Will Be Useful to Cells

Since each of the 20 amino acids is chemically distinct and each can, in principle, occur at any position in a protein chain, there are 20 × 20 × 20 × 20 = 160,000 different possible polypeptide chains four amino acids long, or 20n different possible polypeptide chains n amino acids long. For a typical protein length of about 300 amino acids, a cell could theoretically make more than 10390 (20300) different polypeptide chains. This is such an enormous number that to produce just one molecule of each kind would require many more atoms than exist in the universe.

Only a very small fraction of this vast set of conceivable polypeptide chains would adopt a stable three-dimensional conformation—by some estimates, less than one in a billion. And yet the majority of proteins present in cells do adopt unique and stable conformations. How is this possible? The answer lies in natural selection. A protein with an unpredictably variable structure and biochemical activity is unlikely to help the survival of a cell that contains it. Such proteins would therefore have been eliminated by natural selection through the enormously long trial-and-error process that underlies biological evolution.

Because evolution has selected for protein function in living organisms, the amino acid sequence of most present-day proteins is such that a single conformation is stable. In addition, this conformation has its chemical properties finely tuned to enable the protein to perform a particular catalytic or structural function in the cell. Proteins are so precisely built that the change of even a few atoms in one amino acid can sometimes disrupt the structure of the whole molecule so severely that all function is lost. And, as discussed later in this chapter, when certain rare protein misfolding accidents occur, the results can be disastrous for the organisms that contain them.

Proteins Can Be Classified into Many Families

Once a protein had evolved that folded up into a stable conformation with useful properties, its structure could be modified during evolution to enable it to perform new functions. This process has been greatly accelerated by genetic mechanisms that occasionally duplicate genes, allowing one gene copy to evolve independently to perform a new function (discussed in Chapter 4). This type of event has occurred very often in the past; as a result, many present-day proteins can be grouped into protein families, each family member having an amino acid sequence and a three-dimensional conformation that resemble those of the other family members.

Consider, for example, the serine proteases, a large family of protein-cleaving (proteolytic) enzymes that includes the digestive enzymes chymotrypsin, trypsin, and elastase, and several proteases involved in blood clotting. When the protease portions of any two of these enzymes are compared, parts of their amino acid sequences are found to match. The similarity of their three-dimensional conformations is even more striking: most of the detailed twists and turns in their polypeptide chains, which are several hundred amino acids long, are virtually identical (Figure 3–12). The many different serine proteases nevertheless have distinct enzymatic activities, each cleaving different proteins or the peptide bonds between different types of amino acids. Each therefore performs a distinct function in an organism.

The story we have told for the serine proteases could be repeated for hundreds of other protein families. In general, the structure of the different members of a protein family has been more highly conserved than has the amino acid sequence. In many cases, the amino acid sequences have diverged so far that we cannot be certain of a family relationship between two proteins without determining their three-dimensional structures. The yeast α2 protein and the Drosophila engrailed protein, for example, are both gene regulatory proteins in the homeodomain family (discussed in Chapter 7).

Because they are identical in only 17 of their 60 amino acid residues, their relationship became certain only by comparing their three-dimensional structures (Figure 3–13). Many similar examples show that two proteins with more than 25% identity in their amino acid sequences usually share the same overall structure.

The various members of a large protein family often have distinct functions. Some of the amino acid changes that make family members different were no doubt selected in the course of evolution because they resulted in useful changes in biological activity, giving the individual family members the different functional properties they have today. But many other amino acid changes are effectively “neutral,” having neither a beneficial nor a damaging effect on the basic structure and function of the protein. In addition, since mutation is a random process, there must also have been many deleterious changes that altered the three-dimensional structure of these proteins sufficiently to harm them. Such faulty proteins would have been lost whenever the individual organisms making them were at enough of a disadvantage to be eliminated by natural selection.

Protein families are readily recognized when the genome of any organism is sequenced; for example, the determination of the DNA sequence for the entire human genome has revealed that we contain about 21,000 protein-coding genes. (Note, however, that as a result of alternative RNA splicing, human cells can produce much more than 21,000 different proteins, as will be explained in Chapter 6.) Through sequence comparisons, we can assign the products of at least 40% of our protein-coding genes to known protein structures, belonging to more than 500 different protein families. Most of the proteins in each family have evolved to perform somewhat different functions, as for the enzymes elastase and chymotrypsin illustrated previously in Figure 3–12. As explained in Chapter 1 (see Figure 1–21), these are sometimes called paralogs to distinguish them from the many corresponding proteins in different organisms (orthologs, such as mouse and human elastase).

As described in Chapter 8, because of the powerful techniques of x-ray crystallography and nuclear magnetic resonance (NMR), we now know the three-dimensional shapes, or conformations, of more than 100,000 proteins. By carefully comparing the conformations of these proteins, structural biologists (that is, experts on the structure of biological molecules) have concluded that there are a limited number of ways in which protein domains fold up in nature—maybe as few as 2000, if we consider all organisms. For most of these so-called protein folds, representative structures have been determined.

The present database of known protein sequences contains more than twenty million entries, and it is growing very rapidly as more and more genomes are sequenced—revealing huge numbers of new genes that encode proteins. The encoded polypeptides range widely in size, from 6 amino acids to a gigantic protein of 33,000 amino acids. Protein comparisons are important because related structures often imply related functions. Many years of experimentation can be saved by discovering that a new protein has an amino acid sequence similarity with a protein of known function. Such sequence relationships, for example, first indicated that certain genes that cause mammalian cells to become cancerous encode protein kinases (discussed in Chapter 20).

Some Protein Domains Are Found in Many Different Proteins

As previously stated, most proteins are composed of a series of protein domains, in which different regions of the polypeptide chain fold independently to form compact structures. Such multidomain proteins are believed to have originated from the accidental joining of the DNA sequences that encode each domain, creating a new gene. In an evolutionary process called domain shuffling, many large proteins have evolved through the joining of preexisting domains in new combinations (Figure 3–14). Novel binding surfaces have often been created at the juxtaposition of domains, and many of the functional sites where proteins bind to small molecules are found to be located there.

A subset of protein domains has been especially mobile during evolution; these seem to have particularly versatile structures and are sometimes referred to as protein modules. The structure of one, the SH2 domain, was illustrated in Figure 3–6. Three other abundant protein domains are illustrated in Figure 3–15.

Each of the domains shown has a stable core structure formed from strands of β sheets, from which less-ordered loops of polypeptide chain protrude. The loops are ideally situated to form binding sites for other molecules, as most clearly demonstrated for the immunoglobulin fold, which forms the basis for antibody molecules. Such β-sheet-based domains may have achieved their evolutionary success because they provide a convenient framework for the generation of new binding sites for ligands, requiring only small changes to their protruding loops (see Figure 3–42).

A second feature of these protein domains that explains their utility is the ease with which they can be integrated into other proteins. Two of the three domains illustrated in Figure 3–15 have their N- and C-terminal ends at opposite poles of the domain.

When the DNA encoding such a domain undergoes tandem duplication, which is not unusual in the evolution of genomes (discussed in Chapter 4), the duplicated domains with this “in-line” arrangement can be readily linked in series to form extended structures—either with themselves or with other in-line domains (Figure 3–16). Stiff extended structures composed of a series of domains are especially common in extracellular matrix molecules and in the extracellular portions of cell-surface receptor proteins. Other frequently used domains, including the kringle domain illustrated in Figure 3–15 and the SH2 domain, are of a “plug-in” type, with their N- and C-termini close together. After genomic rearrangements, such domains are usually accommodated as an insertion into a loop region of a second protein.

A comparison of the relative frequency of domain utilization in different eukaryotes reveals that, for many common domains, such as protein kinases, this frequency is similar in organisms as diverse as yeast, plants, worms, flies, and humans. But there are some notable exceptions, such as the Major Histocompatibility Complex (MHC) antigen-recognition domain (see Figure 24–36) that is present in 57 copies in humans, but absent in the other four organisms just mentioned. Domains such as these have specialized functions that are not shared with the other eukaryotes; they are assumed to have been strongly selected for during recent evolution to produce the multiple copies observed. Similarly, the SH2 domain shows an unusual increase in its numbers in higher eukaryotes; such domains might be assumed to be especially useful for multicellularity.

Certain Pairs of Domains Are Found Together in Many Proteins

We can construct a large table displaying domain usage for each organism whose genome sequence is known. For example, the human genome contains the DNA sequences for about 1000 immunoglobulin domains, 500 protein kinase domains, 250 DNA-binding homeodomains, 300 SH3 domains, and 120 SH2 domains. In addition, we find that more than two-thirds of all proteins consist of two or more domains, and that the same pairs of domains occur repeatedly in the same relative arrangement in a protein. Although half of all domain families are common to archaea, bacteria, and eukaryotes, only about 5% of the two-domain combinations are similarly shared. This pattern suggests that most proteins containing especially useful two-domain combinations arose through domain shuffling relatively late in evolution.

The Human Genome Encodes a Complex Set of Proteins, Revealing That Much Remains Unknown

The result of sequencing the human genome has been surprising, because it reveals that our chromosomes contain only about 21,000 protein-coding genes. Based on this number alone, we would appear to be no more complex than the tiny mustard weed, Arabidopsis, and only about 1.3-fold more complex than a nematode worm. The genome sequences also reveal that vertebrates have inherited nearly all of their protein domains from invertebrates—with only 7% of identified human domains being vertebrate-specific.

Each of our proteins is on average more complicated, however (Figure 3–17). Domain shuffling during vertebrate evolution has given rise to many novel combinations of protein domains, with the result that there are nearly twice as many combinations of domains found in human proteins as in a worm or a fly. Thus, for example, the trypsinlike serine protease domain is linked to at least 18 other types of protein domains in human proteins, whereas it is found covalently joined to only 5 different domains in the worm. This extra variety in our proteins greatly increases the range of protein–protein interactions possible (see Figure 3–79), but how it contributes to making us human is not known.

The complexity of living organisms is staggering, and it is quite sobering to note that we currently lack even the tiniest hint of what the function might be for more than 10,000 of the proteins that have thus far been identified through examining the human genome. There are certainly enormous challenges ahead for the next generation of cell biologists, with no shortage of fascinating mysteries to solve.

Larger Protein Molecules Often Contain More Than One Polypeptide Chain

The same weak noncovalent bonds that enable a protein chain to fold into a specific conformation also allow proteins to bind to each other to produce larger structures in the cell. Any region of a protein’s surface that can interact with another molecule through sets of noncovalent bonds is called a binding site. A protein can contain binding sites for various large and small molecules. If a binding site recognizes the surface of a second protein, the tight binding of two folded polypeptide chains at this site creates a larger protein molecule with a precisely defined geometry. Each polypeptide chain in such a protein is called a protein subunit.

In the simplest case, two identical folded polypeptide chains bind to each other in a “head-to-head” arrangement, forming a symmetric complex of two protein subunits (a dimer) held together by interactions between two identical binding sites. The Cro repressor protein—a viral gene regulatory protein that binds to DNA to turn specific viral genes off in an infected bacterial cell—provides an example (Figure 3–18).

Cells contain many other types of symmetric protein complexes, formed from multiple copies of a single polypeptide chain (for example, see Figure 3–20 below).

Many of the proteins in cells contain two or more types of polypeptide chains. Hemoglobin, the protein that carries oxygen in red blood cells, contains two identical α-globin subunits and two identical β-globin subunits, symmetrically arranged (Figure 3–19). Such multisubunit proteins are very common in cells, and they can be very large (Movie 3.6).

Some Globular Proteins Form Long Helical Filaments

Most of the proteins that we have discussed so far are globular proteins, in which the polypeptide chain folds up into a compact shape like a ball with an irregular surface. Some of these protein molecules can nevertheless assemble to form filaments that may span the entire length of a cell. Most simply, a long chain of identical protein molecules can be constructed if each molecule has a binding site complementary to another region of the surface of the same molecule (Figure 3–20). An actin filament, for example, is a long helical structure produced from many molecules of the protein actin (Figure 3–21). Actin is a globular protein that is very abundant in eukaryotic cells, where it forms one of the major filament systems of the cytoskeleton (discussed in Chapter 16).

We will encounter many helical structures in this book. Why is a helix such a common structure in biology? As we have seen, biological structures are often formed by linking similar subunits into long, repetitive chains. If all the subunits are identical, the neighboring subunits in the chain can often fit together in only one way, adjusting their relative positions to minimize the free energy of the contact between them. As a result, each subunit is positioned in exactly the same way in relation to the next, so that subunit 3 fits onto subunit 2 in the same way that subunit 2 fits onto subunit 1, and so on. Because it is very rare for subunits to join up in a straight line, this arrangement generally results in a helix—a regular structure that resembles a spiral staircase, as illustrated in Figure 3–22. Depending on the twist of the staircase, a helix is said to be either right-handed or left-handed (see Figure 3–22E). Handedness is not affected by turning the helix upside down, but it is reversed if the helix is reflected in the mirror.

The observation that helices occur commonly in biological structures holds true whether the subunits are small molecules linked together by covalent bonds (for example, the amino acids in an α helix) or large protein molecules that are linked by noncovalent forces (for example, the actin molecules in actin filaments). This is not surprising.

A helix is an unexceptional structure, and it is generated simply by placing many similar subunits next to each other, each in the same strictly repeated relationship to the one before—that is, with a fixed rotation followed by a fixed translation along the helix axis, as in a spiral staircase.

Many Protein Molecules Have Elongated, Fibrous Shapes

Enzymes tend to be globular proteins: even though many are large and complicated, with multiple subunits, most have an overall rounded shape. In Figure 3–21, we saw that a globular protein can also associate to form long filaments. But there are also functions that require each individual protein molecule to span a large distance. These proteins generally have a relatively simple, elongated three-dimensional structure and are commonly referred to as fibrous proteins.

One large family of intracellular fibrous proteins consists of α-keratin, introduced when we presented the α helix, and its relatives. Keratin filaments are extremely stable and are the main component in long-lived structures such as hair, horn, and nails. An α-keratin molecule is a dimer of two identical subunits, with the long α helices of each subunit forming a coiled-coil (see Figure 3–9). The coiled-coil regions are capped at each end by globular domains containing binding sites. This enables this class of protein to assemble into ropelike intermediate filaments—an important component of the cytoskeleton that creates the cell’s internal structural framework (see Figure 16–67).

Fibrous proteins are especially abundant outside the cell, where they are a main component of the gel-like extracellular matrix that helps to bind collections of cells together to form tissues. Cells secrete extracellular matrix proteins into their surroundings, where they often assemble into sheets or long fibrils. Collagen is the most abundant of these proteins in animal tissues. A collagen molecule consists of three long polypeptide chains, each containing the nonpolar amino acid glycine at every third position. This regular structure allows the chains to wind around one another to generate a long regular triple helix (Figure 3–23A). Many collagen molecules then bind to one another side-by-side and end-toend to create long overlapping arrays—thereby generating the extremely tough collagen fibrils that give connective tissues their tensile strength, as described in Chapter 19.

Proteins Contain a Surprisingly Large Amount of Intrinsically Disordered Polypeptide Chain

It has been well known for a long time that, in complete contrast to collagen, another abundant protein in the extracellular matrix, elastin, is formed as a highly disordered polypeptide.

This disorder is essential for elastin’s function. Its relatively loose and unstructured polypeptide chains are covalently cross-linked to produce a rubberlike, elastic meshwork that can be reversibly pulled from one conformation to another, as illustrated in Figure 3–23B. The elastic fibers that result enable skin and other tissues, such as arteries and lungs, to stretch and recoil without tearing.

Intrinsically disordered regions of proteins are frequent in nature, and they have important functions in the interior of cells. As we have already seen, proteins often have loops of polypeptide chain that protrude from the core region of a protein domain to bind other molecules. Some of these loops remain largely unstructured until they bind to a target molecule, adopting a specific folded conformation only when this other molecule is bound. Many proteins were also known to have intrinsically disordered tails at one or the other end of a structured domain (see, for example, the histones in Figure 4–24). But the extent of such disordered structure only became clear when genomes were sequenced. This allowed bioinformatic methods to be used to analyze the amino acid sequences that genes encode, searching for disordered regions based on their unusually low hydrophobicity and relatively high net charge. Combining these results with other data, it is now thought that perhaps a quarter of all eukaryotic proteins can adopt structures that are mostly disordered, fluctuating rapidly between many different conformations. Many such intrinsically disordered regions contain repeated sequences of amino acids. What do these disordered regions do?

Some known functions are illustrated in Figure 3–24. One predominant function is to form specific binding sites for other protein molecules that are of high specificity, but readily altered by protein phosphorylation, protein dephosphorylation, or any of the other covalent modifications that are triggered by cell signaling events (Figure 3–24A and B). We shall see, for example, that the eukaryotic RNA polymerase enzyme that produces mRNAs contains a long, unstructured C-terminal tail that is covalently modified as its RNA synthesis proceeds, thereby attracting specific other proteins to the transcription complex at different times (see Figure 6–22). And this unstructured tail interacts with a different type of low complexity domain when the RNA polymerase is recruited to the specific sites on the DNA where it begins synthesis.

As illustrated in Figure 3–24C, an unstructured region can also serve as a “tether” to hold two protein domains in close proximity to facilitate their interaction. For example, it is this tethering function that allows substrates to move between active sites in large multienzyme complexes (see Figure 3–54).

A similar tethering function allows large scaffold proteins with multiple protein-binding sites to concentrate sets of interacting proteins, both increasing reaction rates and confining their reaction to a particular site in a cell (see Figure 3–78).

Like elastin, other proteins have a function that directly requires that they remain largely unstructured. Thus, large numbers of disordered protein chains in close proximity can create micro-regions of gel-like consistency inside the cell that restrict diffusion. For example, the abundant nucleoporins that coat the inner surface of the nuclear pore complex form a random coil meshwork (Figure 3–24) that is critical for selective nuclear transport (see Figure 12–8).

Covalent Cross-Linkages Stabilize Extracellular Proteins

Many protein molecules are either attached to the outside of a cell’s plasma membrane or secreted as part of the extracellular matrix. All such proteins are directly exposed to extracellular conditions. To help maintain their structures, the polypeptide chains in such proteins are often stabilized by covalent cross-linkages. These linkages can either tie together two amino acids in the same protein, or connect different polypeptide chains in a multisubunit protein. Although many other types exist, the most common cross-linkages in proteins are covalent sulfur–sulfur bonds. These disulfide bonds (also called S–S bonds) form as cells prepare newly synthesized proteins for export. As described in Chapter

12, their formation is catalyzed in the endoplasmic reticulum by an enzyme that links together two pairs of –SH groups of cysteine side chains that are adjacent in the folded protein (Figure 3–25). Disulfide bonds do not change the conformation of a protein but instead act as atomic staples to reinforce its most favored conformation. For example, lysozyme—an enzyme in tears that dissolves bacterial cell walls—retains its antibacterial activity for a long time because it is stabilized by such cross-linkages.

Disulfide bonds generally fail to form in the cytosol, where a high concentration of reducing agents converts S–S bonds back to cysteine –SH groups. Apparently, proteins do not require this type of reinforcement in the relatively mild environment inside the cell.

Protein Molecules Often Serve as Subunits for the Assembly of Large Structures

The same principles that enable a protein molecule to associate with itself to form rings or a long filament also operate to generate much larger structures formed from a set of different macromolecules, such as enzyme complexes, ribosomes, viruses, and membranes. These large objects are not made as single, giant, covalently linked molecules. Instead they are formed by the noncovalent assembly of many separately manufactured molecules, which serve as the subunits of the final structure.

The use of smaller subunits to build larger structures has several advantages:

1. A large structure built from one or a few repeating smaller subunits requires only a small amount of genetic information.
2. Both assembly and disassembly can be readily controlled reversible processes, because the subunits associate through multiple bonds of relatively low energy.
3. Errors in the synthesis of the structure can be more easily avoided, since correction mechanisms can operate during the course of assembly to exclude malformed subunits.

Some protein subunits assemble into flat sheets in which the subunits are arranged in hexagonal patterns. Specialized membrane proteins are sometimes arranged this way in lipid bilayers. With a slight change in the geometry of the individual subunits, a hexagonal sheet can be converted into a tube (Figure 3–26) or, with more changes, into a hollow sphere. Protein tubes and spheres that bind specific RNA and DNA molecules in their interior form the coats of viruses.

The formation of closed structures, such as rings, tubes, or spheres, provides additional stability because it increases the number of bonds between the protein subunits. Moreover, because such a structure is created by mutually dependent, cooperative interactions between subunits, a relatively small change that affects each subunit individually can cause the structure to assemble or disassemble. These principles are dramatically illustrated in the protein coat or capsid of many simple viruses, which takes the form of a hollow sphere based on an icosahedron (Figure 3–27). Capsids are often made of hundreds of identical protein subunits that enclose and protect the viral nucleic acid (Figure 3–28). The protein in such a capsid must have a particularly adaptable structure: not only must it make several different kinds of contacts to create the sphere, it must also change this arrangement to let the nucleic acid out to initiate viral replication once the virus has entered a cell.

Many Structures in Cells Are Capable of Self-Assembly

The information for forming many of the complex assemblies of macromolecules in cells must be contained in the subunits themselves, because purified subunits can spontaneously assemble into the final structure under the appropriate conditions. The first large macromolecular aggregate shown to be capable of self-assembly from its component parts was tobacco mosaic virus (TMV). This virus is a long rod in which a cylinder of protein is arranged around a helical RNA core (Figure 3–29). If the dissociated RNA and protein subunits are mixed together in solution, they recombine to form fully active viral particles. The assembly process is unexpectedly complex and includes the formation of double rings of protein, which serve as intermediates that add to the growing viral coat.

Another complex macromolecular aggregate that can reassemble from its component parts is the bacterial ribosome. This structure is composed of about 55 different protein molecules and 3 different rRNA molecules. Incubating a mixture of the individual components under appropriate conditions in a test tube causes them to spontaneously re-form the original structure. Most importantly, such reconstituted ribosomes are able to catalyze protein synthesis. As might be expected, the reassembly of ribosomes follows a specific pathway: after certain proteins have bound to the RNA, this complex is then recognized by other proteins, and so on, until the structure is complete.

It is still not clear how some of the more elaborate self-assembly processes are regulated. Many structures in the cell, for example, seem to have a precisely defined length that is many times greater than that of their component macromolecules. How such length determination is achieved is in many cases a mystery. In the simplest case, a long core protein or other macromolecule provides a scaffold that determines the extent of the final assembly. This is the mechanism that determines the length of the TMV particle, where the RNA chain provides the core. Similarly, a core protein interacting with actin is thought to determine the length of the thin filaments in muscle.

Assembly Factors Often Aid the Formation of Complex Biological Structures

Not all cellular structures held together by noncovalent bonds self-assemble. A cilium, or a myofibril of a muscle cell, for example, cannot form spontaneously from a solution of its component macromolecules. In these cases, part of the assembly information is provided by special enzymes and other proteins that perform the function of templates, serving as assembly factors that guide construction but take no part in the final assembled structure.

Even relatively simple structures may lack some of the ingredients necessary for their own assembly. In the formation of certain bacterial viruses, for example, the head, which is composed of many copies of a single protein subunit, is assembled on a temporary scaffold composed of a second protein that is produced by the virus. Because the second protein is absent from the final viral particle, the head structure cannot spontaneously reassemble once it has been taken apart. Other examples are known in which proteolytic cleavage is an essential and irreversible step in the normal assembly process. This is even the case for some small protein assemblies, including the structural protein collagen and the hormone insulin (Figure 3–30). From these relatively simple examples, it seems certain that the assembly of a structure as complex as a cilium will involve a temporal and spatial ordering that is imparted by numerous other components.

Amyloid Fibrils Can Form from Many Proteins

A special class of protein structures, utilized for some normal cell functions, can also contribute to human diseases when not controlled. These are self-propagating, stable β-sheet aggregates called amyloid fibrils. These fibrils are built from a series of identical polypeptide chains that become layered one over the other to create a continuous stack of β sheets, with the β strands oriented perpendicular to the fibril axis to form a cross-beta filament (Figure 3–31). Typically, hundreds of monomers will aggregate to form an unbranched fibrous structure that is several micrometers long and 5 to 15 nm in width. A surprisingly large fraction of proteins have the potential to form such structures, because the short segment of the polypeptide chain that forms the spine of the fibril can have a variety of different sequences and follow one of several different paths (Figure 3–32). However, very few proteins will actually form this structure inside cells.

In normal humans, the quality control mechanisms governing proteins gradually decline with age, occasionally permitting normal proteins to form pathological aggregates. The protein aggregates may be released from dead cells and accumulate as amyloid in the extracellular matrix. In extreme cases, the accumulation of such amyloid fibrils in the cell interior can kill the cells and damage tissues. Because the brain is composed of a highly organized collection of nerve cells that cannot regenerate, the brain is especially vulnerable to this sort of cumulative damage. Thus, although amyloid fibrils may form in different tissues, and are known to cause pathologies in several sites in the body, the most severe amyloid pathologies are neurodegenerative diseases. For example, the abnormal formation of highly stable amyloid fibrils is thought to play a central causative role in both Alzheimer’s and Parkinson’s diseases.

Prion diseases are a special type of these pathologies. They have attained special notoriety because, unlike Parkinson’s or Alzheimer’s, prion diseases can spread from one organism to another, providing that the second organism eats a tissue containing the protein aggregate. A set of closely related diseases—scrapie in sheep, Creutzfeldt–Jakob disease (CJD) in humans, Kuru in humans, and bovine spongiform encephalopathy (BSE) in cattle—are caused by a misfolded, aggregated form of a particular protein called PrP (for prion protein). PrP is normally located on the outer surface of the plasma membrane, most prominently in neurons, and it has the unfortunate property of forming amyloid fibrils that are “infectious” because they convert normally folded molecules of PrP to the same pathological form (Figure 3–33).

This property creates a positive feedback loop that propagates the abnormal form of PrP, called PrP*, and allows the pathological conformation to spread rapidly from cell to cell in the brain, eventually causing death. It can be dangerous to eat the tissues of animals that contain PrP*, as witnessed by the spread of BSE (commonly referred to as “mad cow disease”) from cattle to humans. Fortunately, in the absence of PrP*, PrP is extraordinarily difficult to convert to its abnormal form.

A closely related “protein-only inheritance” has been observed in yeast cells. The ability to study infectious proteins in yeast has clarified another remarkable feature of prions. These protein molecules can form several distinctively different types of amyloid fibrils from the same polypeptide chain. Moreover, each type of aggregate can be infectious, forcing normal protein molecules to adopt the same type of abnormal structure. Thus, several different “strains” of infectious particles can arise from the same polypeptide chain.

Amyloid Structures Can Perform Useful Functions in Cells

Amyloid fibrils were initially studied because they cause disease. But the same type of structure is now known to be exploited by cells for useful purposes. Eukaryotic cells, for example, store many different peptide and protein hormones that they will secrete in specialized “secretory granules,” which package a high concentration of their cargo in dense cores with a regular structure (see Figure 13–65). We now know that these structured cores consist of amyloid fibrils, which in this case have a structure that causes them to dissolve to release soluble cargo after being secreted by exocytosis to the cell exterior (Figure 3–34A). Many bacteria use the amyloid structure in a very different way, secreting proteins that form long amyloid fibrils projecting from the cell exterior that help to bind bacterial neighbors into biofilms (Figure 3–34B). Because these biofilms help bacteria to survive in adverse environments (including in humans treated with antibiotics), new drugs that specifically disrupt the fibrous networks formed by bacterial amyloids have promise for treating human infections.

Many Proteins Contain Low-complexity Domains that Can Form “Reversible Amyloids”

Until recently, those amyloids with useful functions were thought to be either confined to the interior of specialized vesicles or expressed on the exterior of cells, as in Figure 3–34. However, new experiments reveal that a large set of low complexity domains can form amyloid fibers that have functional roles in both the cell nucleus and the cell cytoplasm.

These domains are normally unstructured and consist of stretches of amino acid sequence that can span hundreds of amino acids, while containing only a small subset of the 20 different amino acids. In contrast to the disease-associated amyloid in Figure 3–33, these newly discovered structures are held together by weaker noncovalent bonds and readily dissociate in response to signals—hence their name reversible amyloids.

Many proteins with such domains also contain a different set of domains that bind to specific other protein or RNA molecules. Thus, their controlled aggregation in the cell can form a hydrogel that pulls these and other molecules into punctate structures called intracellular bodies, or granules. Specific mRNAs can be sequestered in such granules, where they are stored until made available by a controlled disassembly of the core amyloid structure that holds them together.

Consider the FUS protein, an essential nuclear protein with roles in the transcription, processing, and transport of specific mRNA molecules. Over 80 percent of its C-terminal domain of two hundred amino acids is composed of only four amino acids: glycine, serine, glutamine, and tyrosine. This low complexity domain is attached to several other domains that bind to RNA molecules. At high enough concentrations in a test tube, it forms a hydrogel that will associate with either itself or with the low complexity domains from other proteins. As illustrated by the experiment in Figure 3–35, although different low complexity domains bind to each other, homotypic interactions appear to be of greatest affinity (thus, the FUS low complexity domain binds most tightly to itself). Further experiments reveal that that both the homotypic and the heterotypic bindings are mediated through a β-sheet core structure forming amyloid fibrils, and that these structures bind to other types of repeat sequences in the manner indicated in Figure 3–36.

Many of these interactions appear to be controlled by the phosphorylation of serine side chains in the one or both of the interacting partners. However, a great deal remains to be learned concerning these newly discovered structures and the varied roles that they play in the cell biology of eukaryotic cells.

Summary

A protein molecule’s amino acid sequence determines its three-dimensional conformation. Noncovalent interactions between different parts of the polypeptide chain stabilize its folded structure. The amino acids with hydrophobic side chains tend to cluster in the interior of the molecule, and local hydrogen-bond interactions between neighboring peptide bonds give rise to α helices and β sheets.

Regions of amino acid sequence known as domains are the modular units from which many proteins are constructed. Such domains generally contain 40–350 amino acids, often folded into a globular shape. Small proteins typically consist of only a single domain, while large proteins are formed from multiple domains linked together by various lengths of polypeptide chain, some of which can be relatively disordered. As proteins have evolved, domains have been modified and combined with other domains to construct large numbers of new proteins.

Proteins are brought together into larger structures by the same noncovalent forces that determine protein folding. Proteins with binding sites for their own surface can assemble into dimers, closed rings, spherical shells, or helical polymers. The amyloid fibril is a long unbranched structure assembled through a repeating aggregate of β sheets. Although some mixtures of proteins and nucleic acids can assemble spontaneously into complex structures in a test tube, not all structures in the cell are capable of spontaneous reassembly after they have been dissociated into their component parts, because many biological assembly processes involve assembly factors that are not present in the final structure.

PROTEIN FUNCTION

We have seen that each type of protein consists of a precise sequence of amino acids that allows it to fold up into a particular three-dimensional shape, or conformation. But proteins are not rigid lumps of material. They often have precisely engineered moving parts whose mechanical actions are coupled to chemical events. It is this coupling of chemistry and movement that gives proteins the extraordinary capabilities that underlie the dynamic processes in living cells.

In this section, we explain how proteins bind to other selected molecules and how a protein's activity depends on such binding. We show that the ability to bind to other molecules enables proteins to act as catalysts, signal receptors, switches, motors, or tiny pumps. The examples we discuss in this chapter by no means exhaust the vast functional repertoire of proteins. You will encounter the specialized functions of many other proteins elsewhere in this book, based on similar principles.

All Proteins Bind to Other Molecules

A protein molecule's physical interaction with other molecules determines its biological properties. Thus, antibodies attach to viruses or bacteria to mark them for destruction, the enzyme hexokinase binds glucose and ATP so as to catalyze a reaction between them, actin molecules bind to each other to assemble into actin filaments, and so on. Indeed, all proteins stick, or bind, to other molecules.

In some cases, this binding is very tight; in others it is weak and short-lived. But the binding always shows great specificity, in the sense that each protein molecule can usually bind just one or a few molecules out of the many thousands of different types it encounters. The substance that is bound by the protein—whether it is an ion, a small molecule, or a macromolecule such as another protein—is referred to as a ligand for that protein (from the Latin word ligare, meaning “to bind”).

The ability of a protein to bind selectively and with high affinity to a ligand depends on the formation of a set of weak noncovalent bonds—hydrogen bonds, electrostatic attractions, and van der Waals attractions—plus favorable hydrophobic interactions (see Panel 2–3, pp. 94–95). Because each individual bond is weak, effective binding occurs only when many of these bonds form simultaneously. Such binding is possible only if the surface contours of the ligand molecule fit very closely to the protein, matching it like a hand in a glove (Figure 3–37).

The region of a protein that associates with a ligand, known as the ligand's binding site, usually consists of a cavity in the protein surface formed by a particular arrangement of amino acids. These amino acids can belong to different portions of the polypeptide chain that are brought together when the protein folds (Figure 3–38). Separate regions of the protein surface generally provide binding sites for different ligands, allowing the protein's activity to be regulated, as we shall see later. And other parts of the protein act as a handle to position the protein in the cell—an example is the SH2 domain discussed previously, which often moves a protein containing it to particular intracellular sites in response to signals.

Although the atoms buried in the interior of the protein have no direct contact with the ligand, they form the framework that gives the surface its contours and its chemical and mechanical properties. Even small changes to the amino acids in the interior of a protein molecule can change its three-dimensional shape enough to destroy a binding site on the surface.

The Surface Conformation of a Protein Determines Its Chemistry

The impressive chemical capabilities of proteins often require that the chemical groups on their surface interact in ways that enhance the chemical reactivity of one or more amino acid side chains. These interactions fall into two main categories.

First, the interaction of neighboring parts of the polypeptide chain may restrict the access of water molecules to that protein’s ligand-binding sites. Because water molecules readily form hydrogen bonds that can compete with ligands for sites on the protein surface, a ligand will form tighter hydrogen bonds (and electrostatic interactions) with a protein if water molecules are kept away.

It might be hard to imagine a mechanism that would exclude a molecule as small as water from a protein surface without affecting the access of the ligand itself. However, because of the strong tendency of water molecules to form water–water hydrogen bonds, water molecules exist in a large hydrogen-bonded network (see Panel 2–2, pp. 92–93). In effect, a protein can keep a ligand-binding site dry, increasing that site's reactivity, because it is energetically unfavorable for individual water molecules to break away from this network—as they must do to reach into a crevice on a protein’s surface.

Second, the clustering of neighboring polar amino acid side chains can alter their reactivity. If protein folding forces together a number of negatively charged side chains against their mutual repulsion, for example, the affinity of the site for a positively charged ion is greatly increased. In addition, when amino acid side chains interact with one another through hydrogen bonds, normally unreactive groups (such as the –CH2OH on the serine shown in Figure 3–39) can become reactive, enabling them to be used to make or break selected covalent bonds.

The surface of each protein molecule therefore has a unique chemical reactivity that depends not only on which amino acid side chains are exposed, but also on their exact orientation relative to one another. For this reason, two slightly different conformations of the same protein molecule can differ greatly in their chemistry.

Sequence Comparisons Between Protein Family Members Highlight Crucial Ligand-Binding Sites

As we have described previously, genome sequences allow us to group many of the domains in proteins into families that show clear evidence of their evolution from a common ancestor. The three-dimensional structures of members of the same domain family are remarkably similar. For example, even when the amino acid sequence identity falls to 25%, the backbone atoms in a domain can follow a common protein fold within 0.2 nanometers (2 Å).

We can use a method called evolutionary tracing to identify those sites in a protein domain that are the most crucial to the domain’s function. Those sites that bind to other molecules are the most likely to be maintained, unchanged as organisms evolve. Thus, in this method, those amino acids that are unchanged, or nearly unchanged, in all of the known protein family members are mapped onto a model of the three-dimensional structure of one family member. When this is done, the most invariant positions often form one or more clusters on the protein surface, as illustrated in Figure 3–40A for the SH2 domain described previously (see Figure 3–6). These clusters generally correspond to ligand-binding sites.

The SH2 domain functions to link two proteins together. It binds the protein containing it to a second protein that contains a phosphorylated tyrosine side chain in a specific amino acid sequence context, as shown in Figure 3–40B. The amino acids located at the binding site for the phosphorylated polypeptide have been the slowest to change during the long evolutionary process that produced the large SH2 family of peptide recognition domains. Mutation is a random process; survival is not. Thus, natural selection (random mutation followed by nonrandom survival) produces the sequence conservation by preferentially eliminating organisms whose SH2 domains become altered in a way that inactivates the SH2 binding site, destroying SH2 function.

Genome sequencing has revealed huge numbers of proteins whose functions are unknown. Once a three-dimensional structure has been determined for one member of a protein family, evolutionary tracing allows biologists to determine binding sites for the members of that family, providing a useful start in deciphering protein function.

Proteins Bind to Other Proteins Through Several Types of Interfaces

Proteins can bind to other proteins in multiple ways. In many cases, a portion of the surface of one protein contacts an extended loop of polypeptide chain (a “string”) on a second protein (Figure 3–41A). Such a surface–string interaction, for example, allows the SH2 domain to recognize a phosphorylated polypeptide loop on a second protein, as just described, and it also enables a protein kinase to recognize the proteins that it will phosphorylate (see below).

A second type of protein–protein interface forms when two α helices, one from each protein, pair together to form a coiled-coil (Figure 3–41B). This type of protein interface is found in several families of gene regulatory proteins, as discussed in Chapter 7.

The most common way for proteins to interact, however, is by the precise matching of one rigid surface with that of another (Figure 3–41C). Such interactions can be very tight, since a large number of weak bonds can form between two surfaces that match well. For the same reason, such surface–surface interactions can be extremely specific, enabling a protein to select just one partner from the many thousands of different proteins found in a cell.

Antibody Binding Sites Are Especially Versatile

All proteins must bind to particular ligands to carry out their various functions. The antibody family is notable for its capacity for tight, highly selective binding (discussed in detail in Chapter 24).

Antibodies, or immunoglobulins, are proteins produced by the immune system in response to foreign molecules, such as those on the surface of an invading microorganism. Each antibody binds tightly to a particular target molecule, thereby either inactivating the target molecule directly or marking it for destruction. An antibody recognizes its target (called an antigen) with remarkable specificity. Because there are potentially billions of different antigens that humans might encounter, we have to be able to produce billions of different antibodies.

Antibodies are Y-shaped molecules with two identical binding sites that are complementary to a small portion of the surface of the antigen molecule. A detailed examination of the antigen-binding sites of antibodies reveals that they are formed from several loops of polypeptide chain that protrude from the ends of a pair of closely juxtaposed protein domains (Figure 3–42). Different antibodies generate an enormous diversity of antigen-binding sites by changing only the length and amino acid sequence of these loops, without altering the basic protein structure.

Loops of this kind are ideal for grasping other molecules. They allow a large number of chemical groups to surround a ligand so that the protein can link to it with many weak bonds. For this reason, loops often form the ligand-binding sites in proteins.

The Equilibrium Constant Measures Binding Strength

Molecules in the cell encounter each other very frequently because of their continual random thermal movements. Colliding molecules with poorly matching surfaces form few noncovalent bonds with one another, and the two molecules dissociate as rapidly as they come together. At the other extreme, when many noncovalent bonds form between two colliding molecules, the association can persist for a very long time (Figure 3–43). Strong interactions occur in cells whenever a biological function requires that molecules remain associated for a long time—for example, when a group of RNA and protein molecules come together to make a subcellular structure such as a ribosome.

We can measure the strength with which any two molecules bind to each other. As an example, consider a population of identical antibody molecules that suddenly encounters a population of ligands diffusing in the fluid surrounding them. At frequent intervals, one of the ligand molecules will bump into the binding site of an antibody and form an antibody–ligand complex. The population of antibody–ligand complexes will therefore increase, but not without limit: over time, a second process, in which individual complexes break apart because of thermally induced motion, will become increasingly important.

Eventually, any population of antibody molecules and ligands will reach a steady state, or equilibrium, in which the number of binding (association) events per second is precisely equal to the number of “unbinding” (dissociation) events (see Figure 2–30).

From the concentrations of the ligand, antibody, and antibody–ligand complex at equilibrium, we can calculate a convenient measure of the strength of binding—the equilibrium constant (K)—(Figure 3–44A). This constant was described in detail in Chapter 2, where its connection to free energy differences was derived (see p. 62). The equilibrium constant for a reaction in which two molecules (A and B) bind to each other to form a complex (AB) has units of liters/mole, and half of the binding sites will be occupied by ligand when that ligand’s concentration (in moles/liter) reaches a value that is equal to 1/K. This equilibrium constant is larger the greater the binding strength, and it is a direct measure of the free-energy difference between the bound and free states (Figure 3–44B). Even a change of a few noncovalent bonds can have a striking effect on a binding interaction, as shown by the example in Figure 3–45. (Note that the equilibrium constant, as defined here, is also known as the association or affinity constant, Ka.)

We have used the case of an antibody binding to its ligand to illustrate the effect of binding strength on the equilibrium state, but the same principles apply to any molecule and its ligand. Many proteins are enzymes, which, as we now discuss, first bind to their ligands and then catalyze the breakage or formation of covalent bonds in these molecules.

Enzymes Are Powerful and Highly Specific Catalysts

Many proteins can perform their function simply by binding to another molecule. An actin molecule, for example, need only associate with other actin molecules to form a filament. There are other proteins, however, for which ligand binding is only a necessary first step in their function. This is the case for the large and very important class of proteins called enzymes. As described in Chapter 2, enzymes are remarkable molecules that cause the chemical transformations that make and break covalent bonds in cells. They bind to one or more ligands, called substrates, and convert them into one or more chemically modified products, doing this over and over again with amazing rapidity. Enzymes speed up reactions, often by a factor of a million or more, without themselves being changed—that is, they act as catalysts that permit cells to make or break covalent bonds in a controlled way. It is the catalysis of organized sets of chemical reactions by enzymes that creates and maintains the cell, making life possible.

We can group enzymes into functional classes that perform similar chemical reactions (Table 3–1). Each type of enzyme within such a class is highly specific, catalyzing only a single type of reaction. Thus, hexokinase adds a phosphate group to D-glucose but ignores its optical isomer L-glucose; the blood-clotting enzyme thrombin cuts one type of blood protein between a particular arginine and its adjacent glycine and nowhere else, and so on. As discussed in detail in Chapter 2, enzymes work in teams, with the product of one enzyme becoming the substrate for the next. The result is an elaborate network of metabolic pathways that provides the cell with energy and generates the many large and small molecules that the cell needs (see Figure 2–63).

Substrate Binding Is the First Step in Enzyme Catalysis

For a protein that catalyzes a chemical reaction (an enzyme), the binding of each substrate molecule to the protein is an essential prelude. In the simplest case, if we denote the enzyme by E, the substrate by S, and the product by P, the basic reaction path is $E + S \rightarrow ES \rightarrow EP \rightarrow E + P$. There is a limit to the amount of substrate that a single enzyme molecule can process in a given time. Although an increase in the concentration of substrate increases the rate at which product is formed, this rate eventually reaches a maximum value (Figure 3–46). At that point the enzyme molecule is saturated with substrate, and the rate of reaction (Vmax) depends only on how rapidly the enzyme can process the substrate molecule. This maximum rate divided by the enzyme concentration is called the turnover number. Turnover numbers are often about 1000 substrate molecules processed per second per enzyme molecule, although turnover numbers between 1 and 10,000 are known.

The other kinetic parameter frequently used to characterize an enzyme is its Km, the concentration of substrate that allows the reaction to proceed at one-half its maximum rate (0.5 Vmax) (see Figure 3–46). A low Km value means that the enzyme reaches its maximum catalytic rate at a low concentration of substrate and generally indicates that the enzyme binds to its substrate very tightly, whereas a high Km value corresponds to weak binding. The methods used to characterize enzymes in this way are explained in Panel 3–2 (pp. 142–143).

Enzymes Speed Reactions by Selectively Stabilizing Transition States

Enzymes achieve extremely high rates of chemical reaction—rates that are far higher than for any synthetic catalysts. There are several reasons for this efficiency. First, when two molecules need to react, the enzyme greatly increases the local concentration of both of these substrate molecules at the catalytic site, holding them in the correct orientation for the reaction that is to follow.

More importantly, however, some of the binding energy contributes directly to the catalysis. Substrate molecules must pass through a series of intermediate states of altered geometry and electron distribution before they form the ultimate products of the reaction. The free energy required to attain the most unstable intermediate state, called the transition state, is known as the activation energy for the reaction, and it is the major determinant of the reaction rate. Enzymes have a much higher affinity for the transition state of the substrate than they have for the stable form.

Because this tight binding greatly lowers the energy of the transition state, the enzyme greatly accelerates a particular reaction by lowering the activation energy that is required (Figure 3–47).

Enzymes Can Use Simultaneous Acid and Base Catalysis

Figure 3–48 compares the spontaneous reaction rates and the corresponding enzyme-catalyzed rates for five enzymes. Rate accelerations range from 109 to

1023. Enzymes not only bind tightly to a transition state, they also contain precisely positioned atoms that alter the electron distributions in the atoms that participate directly in the making and breaking of covalent bonds. Peptide bonds, for example, can be hydrolyzed in the absence of an enzyme by exposing a polypeptide to either a strong acid or a strong base. Enzymes are unique, however, in being able to use acid and base catalysis simultaneously, because the rigid framework of the protein constrains the acidic and basic residues and prevents them from combining with each other, as they would do in solution (Figure 3–49).

The fit between an enzyme and its substrate needs to be precise. A small change introduced by genetic engineering in the active site of an enzyme can therefore have a profound effect. Replacing a glutamic acid with an aspartic acid in one enzyme, for example, shifts the position of the catalytic carboxylate ion by only 1 Å (about the radius of a hydrogen atom); yet this is enough to decrease the activity of the enzyme a thousandfold.

Lysozyme Illustrates How an Enzyme Works

To demonstrate how enzymes catalyze chemical reactions, we examine an enzyme that acts as a natural antibiotic in egg white, saliva, tears, and other secretions. Lysozyme catalyzes the cutting of polysaccharide chains in the cell walls of bacteria. The bacterial cell is under pressure from osmotic forces, and cutting even a small number of these chains causes the cell wall to rupture and the cell to burst. A relatively small and stable protein that can be easily isolated in large quantities, lysozyme was the first enzyme to have its structure worked out in atomic detail by x-ray crystallography (in the mid-1960s).

The reaction that lysozyme catalyzes is a hydrolysis: it adds a molecule of water to a single bond between two adjacent sugar groups in the polysaccharide chain, thereby causing the bond to break (see Figure 2–9). The reaction is energetically favorable because the free energy of the severed polysaccharide chain is lower than the free energy of the intact chain. However, there is an energy barrier to the reaction, and a colliding water molecule can break a bond linking two sugars only if the polysaccharide molecule is distorted into a particular shape—the transition state—in which the atoms around the bond have an altered geometry and electron distribution. Because of this requirement, random collisions must supply a very large activation energy for the reaction to take place. In an aqueous solution at room temperature, the energy of collisions almost never exceeds the activation energy. The pure polysaccharide can therefore remain for years in water without being hydrolyzed to any detectable degree.

This situation changes drastically when the polysaccharide binds to lysozyme. The active site of lysozyme, because its substrate is a polymer, is a long groove that holds six linked sugars at the same time. As soon as the polysaccharide binds to form an enzyme–substrate complex, the enzyme cuts the polysaccharide by adding a water molecule across one of its sugar–sugar bonds. The product chains are then quickly released, freeing the enzyme for further cycles of reaction (Figure 3–50).

An impressive increase in hydrolysis rate is possible because conditions are created in the microenvironment of the lysozyme active site that greatly reduce the activation energy necessary for the hydrolysis to take place. In particular, lysozyme distorts one of the two sugars connected by the bond to be broken from its normal, most stable conformation. The bond to be broken is also held close to two amino acids with acidic side chains (a glutamic acid and an aspartic acid) that participate directly in the reaction. Figure 3–51 shows the three central steps in this enzymatically catalyzed reaction, which occurs millions of times faster than uncatalyzed hydrolysis.

Other enzymes use similar mechanisms to lower activation energies and speed up the reactions they catalyze. In reactions involving two or more reactants, the active site also acts like a template, or mold, that brings the substrates together in the proper orientation for a reaction to occur between them (Figure 3–52A). As we saw for lysozyme, the active site of an enzyme contains precisely positioned atoms that speed up a reaction by using charged groups to alter the distribution of electrons in the substrates (Figure 3–52B).

And as we have also seen, when a substrate binds to an enzyme, bonds in the substrate are often distorted, changing the substrate shape. These changes, along with mechanical forces, drive a substrate toward a particular transition state (Figure 3–52C). Finally, like lysozyme, many enzymes participate intimately in the reaction by transiently forming a covalent bond between the substrate and a side chain of the enzyme. Subsequent steps in the reaction restore the side chain to its original state, so that the enzyme remains unchanged after the reaction (see also Figure 2–48).

Tightly Bound Small Molecules Add Extra Functions to Proteins

Although we have emphasized the versatility of enzymes—and proteins in general—as chains of amino acids that perform remarkable functions, there are many instances in which the amino acids by themselves are not enough. Just as humans employ tools to enhance and extend the capabilities of their hands, enzymes and other proteins often use small nonprotein molecules to perform functions that would be difficult or impossible to do with amino acids alone. Thus, enzymes frequently have a small molecule or metal atom tightly associated with their active site that assists with their catalytic function. Carboxypeptidase, for example, an enzyme that cuts polypeptide chains, carries a tightly bound zinc ion in its active site. During the cleavage of a peptide bond by carboxypeptidase, the zinc ion forms a transient bond with one of the substrate atoms, thereby assisting the hydrolysis reaction. In other enzymes, a small organic molecule serves a similar purpose. Such organic molecules are often referred to as coenzymes. An example is biotin, which is found in enzymes that transfer a carboxylate group (–COO–) from one molecule to another (see Figure 2–40). Biotin participates in these reactions by forming a transient covalent bond to the –COO– group to be transferred, being better suited to this function than any of the amino acids used to make proteins. Because it cannot be synthesized by humans, and must therefore be supplied in small quantities in our diet, biotin is a vitamin. Many other coenzymes are either vitamins or derivatives of vitamins (Table 3–2).

Other proteins also frequently require specific small-molecule adjuncts to function properly. Thus, the signal receptor protein rhodopsin, which is made by the photoreceptor cells in the retina, detects light by means of a small molecule, retinal, embedded in the protein (Figure 3–53A). Retinal, which is derived from vitamin A, changes its shape when it absorbs a photon of light, and this change causes the protein to trigger a cascade of enzymatic reactions that eventually lead to an electrical signal being carried to the brain.

Another example of a protein with a nonprotein portion is hemoglobin (see Figure 3–19). Each molecule of hemoglobin carries four heme groups, ring-shaped molecules each with a single central iron atom (Figure 3–53B). Heme gives hemoglobin (and blood) its red color. By binding reversibly to oxygen gas through its iron atom, heme enables hemoglobin to pick up oxygen in the lungs and release it in the tissues.

Sometimes these small molecules are attached covalently and permanently to their protein, thereby becoming an integral part of the protein molecule itself. We shall see in Chapter 10 that proteins are often anchored to cell membranes through covalently attached lipid molecules. And membrane proteins exposed on the surface of the cell, as well as proteins secreted outside the cell, are often modified by the covalent addition of sugars and oligosaccharides.

Multienzyme Complexes Help to Increase the Rate of Cell Metabolism

The efficiency of enzymes in accelerating chemical reactions is crucial to the maintenance of life. Cells, in effect, must race against the unavoidable processes of decay, which—if left unattended—cause macromolecules to run downhill toward greater and greater disorder. If the rates of desirable reactions were not greater than the rates of competing side reactions, a cell would soon die. We can get some idea of the rate at which cell metabolism proceeds by measuring the rate of ATP utilization. A typical mammalian cell “turns over” (i.e., hydrolyzes and restores by phosphorylation) its entire ATP pool once every 1 or 2 minutes. For each cell, this turnover represents the utilization of roughly 107 molecules of ATP per second (or, for the human body, about 1 gram of ATP every minute).

The rates of reactions in cells are rapid because enzyme catalysis is so effective. Some enzymes have become so efficient that there is no possibility of further useful improvement. The factor that limits the reaction rate is no longer the enzyme’s intrinsic speed of action; rather, it is the frequency with which the enzyme collides with its substrate. Such a reaction is said to be diffusion-limited (see Panel 3–2, pp. 142–143).

The amount of product produced by an enzyme will depend on the concentration of both the enzyme and its substrate. If a sequence of reactions is to occur extremely rapidly, each metabolic intermediate and enzyme involved must be present in high concentration. However, given the enormous number of different reactions performed by a cell, there are limits to the concentrations that can be achieved. In fact, most metabolites are present in micromolar (10–6 M) concentrations, and most enzyme concentrations are much lower. How is it possible, therefore, to maintain very fast metabolic rates?

The answer lies in the spatial organization of cell components. The cell can increase reaction rates without raising substrate concentrations by bringing the various enzymes involved in a reaction sequence together to form a large protein assembly known as a multienzyme complex (Figure 3–54). Because this assembly is organized in a way that allows the product of enzyme A to be passed directly to enzyme B, and so on, diffusion rates need not be limiting, even when the concentrations of the substrates in the cell as a whole are very low. It is perhaps not surprising, therefore, that such enzyme complexes are very common, and they are involved in nearly all aspects of metabolism—including the central genetic processes of DNA, RNA, and protein synthesis. In fact, few enzymes in eukaryotic cells diffuse freely in solution; instead, most seem to have evolved binding sites that concentrate them with other proteins of related function in particular regions of the cell, thereby increasing the rate and efficiency of the reactions that they catalyze (see p. 331).

Eukaryotic cells have yet another way of increasing the rate of metabolic reactions: using their intracellular membrane systems. These membranes can segregate particular substrates and the enzymes that act on them into the same membrane-enclosed compartment, such as the endoplasmic reticulum or the cell nucleus. If, for example, a compartment occupies a total of 10% of the volume of the cell, the concentration of reactants in that compartment may be increased by 10 times compared with a cell with the same number of enzyme and substrate molecules, but no compartmentalization. Reactions limited by the speed of diffusion can thereby be speeded up by a factor of 10.

The Cell Regulates the Catalytic Activities of Its Enzymes

A living cell contains thousands of enzymes, many of which operate at the same time and in the same small volume of the cytosol. By their catalytic action, these enzymes generate a complex web of metabolic pathways, each composed of chains of chemical reactions in which the product of one enzyme becomes the substrate of the next. In this maze of pathways, there are many branch points (nodes) where different enzymes compete for the same substrate. The system is complex (see Figure 2–63), and elaborate controls are required to regulate when and how rapidly each reaction occurs.

Regulation occurs at many levels. At one level, the cell controls how many molecules of each enzyme it makes by regulating the expression of the gene that encodes that enzyme (discussed in Chapter 7). The cell also controls enzymatic activities by confining sets of enzymes to particular subcellular compartments, whether by enclosing them in a distinct membrane-bounded compartment (discussed in Chapters 12 and 14) or by concentrating them on a protein scaffold (see Figure 3–77).

As will be explained later in this chapter, enzymes are also covalently modified to control their activity. The rate of protein destruction by targeted proteolysis represents yet another important regulatory mechanism (see Figure 6–86). But the most general process that adjusts reaction rates operates through a direct, reversible change in the activity of an enzyme in response to the specific small molecules that it binds.

The most common type of control occurs when an enzyme binds a molecule that is not a substrate to a special regulatory site outside the active site, thereby altering the rate at which the enzyme converts its substrates to products. For example, in feedback inhibition, a product produced late in a reaction pathway inhibits an enzyme that acts earlier in the pathway. Thus, whenever large quantities of the final product begin to accumulate, this product binds to the enzyme and slows down its catalytic action, thereby limiting the further entry of substrates into that reaction pathway (Figure 3–55). Where pathways branch or intersect, there are usually multiple points of control by different final products, each of which works to regulate its own synthesis (Figure 3–56). Feedback inhibition can work almost instantaneously, and it is rapidly reversed when the level of the product falls.

Feedback inhibition is negative regulation: it prevents an enzyme from acting. Enzymes can also be subject to positive regulation, in which a regulatory molecule stimulates the enzyme’s activity rather than shutting the enzyme down. Positive regulation occurs when a product in one branch of the metabolic network stimulates the activity of an enzyme in another pathway. As one example, the accumulation of ADP activates several enzymes involved in the oxidation of sugar molecules, thereby stimulating the cell to convert more ADP to ATP.

Allosteric Enzymes Have Two or More Binding Sites That Interact

A striking feature of both positive and negative feedback regulation is that the regulatory molecule often has a shape totally different from the shape of the substrate of the enzyme. This is why the effect on a protein is termed allostery (from the Greek words allos, meaning “other,” and stereos, meaning “solid” or “three-dimensional”). As biologists learned more about feedback regulation, they recognized that the enzymes involved must have at least two different binding sites on their surface—an active site that recognizes the substrates, and a regulatory site that recognizes a regulatory molecule. These two sites must somehow communicate so that the catalytic events at the active site can be influenced by the binding of the regulatory molecule at its separate site on the protein’s surface.

The interaction between separated sites on a protein molecule is now known to depend on a conformational change in the protein: binding at one of the sites causes a shift from one folded shape to a slightly different folded shape. During feedback inhibition, for example, the binding of an inhibitor at one site on the protein causes the protein to shift to a conformation that incapacitates its active site located elsewhere in the protein.

It is thought that most protein molecules are allosteric. They can adopt two or more slightly different conformations, and a shift from one to another caused by the binding of a ligand can alter their activity. This is true not only for enzymes but also for many other proteins, including receptors, structural proteins, and motor proteins. In all instances of allosteric regulation, each conformation of the protein has somewhat different surface contours, and the protein’s binding sites for ligands are altered when the protein changes shape. Moreover, as we discuss next, each ligand will stabilize the conformation that it binds to most strongly, and thus—at high enough concentrations—will tend to “switch” the protein toward the conformation that the ligand prefers.

Two Ligands Whose Binding Sites Are Coupled Must Reciprocally Affect Each Other’s Binding

The effects of ligand binding on a protein follow from a fundamental chemical principle known as linkage. Suppose, for example, that a protein that binds glucose also binds another molecule, X, at a distant site on the protein’s surface. If the binding site for X changes shape as part of the conformational change in the protein induced by glucose binding, the binding sites for X and for glucose are said to be coupled. Whenever two ligands prefer to bind to the same conformation of an allosteric protein, it follows from basic thermodynamic principles that each ligand must increase the affinity of the protein for the other. For example, if the shift of a protein to a conformation that binds glucose best also causes the binding site for X to fit X better, then the protein will bind glucose more tightly when X is present than when X is absent. In other words, X will positively regulate the protein’s binding of glucose (Figure 3–57).

Conversely, linkage operates in a negative way if two ligands prefer to bind to different conformations of the same protein. In this case, the binding of the first ligand discourages the binding of the second ligand. Thus, if a shape change caused by glucose binding decreases the affinity of a protein for molecule X, the binding of X must also decrease the protein’s affinity for glucose (Figure 3–58). The linkage relationship is quantitatively reciprocal, so that, for example, if glucose has a very large effect on the binding of X, X has a very large effect on the binding of glucose.

The relationships shown in Figures 3–57 and 3–58 apply to all proteins, and they underlie all of cell biology. The principle seems so obvious in retrospect that we now take it for granted. But the discovery of linkage in studies of a few enzymes in the 1950s, followed by an extensive analysis of allosteric mechanisms in proteins in the early 1960s, had a revolutionary effect on our understanding of biology. Since molecule X in these examples binds at a site on the enzyme that is distinct from the site where catalysis occurs, it need not have any chemical relationship to the substrate that binds at the active site. Moreover, as we have just seen, for enzymes that are regulated in this way, molecule X can either turn the enzyme on (positive regulation) or turn it off (negative regulation). By such a mechanism, allosteric proteins serve as general switches that, in principle, can allow one molecule in a cell to affect the fate of any other.

Symmetric Protein Assemblies Produce Cooperative Allosteric Transitions

A single-subunit enzyme that is regulated by negative feedback can at most decrease from 90% to about 10% activity in response to a 100-fold increase in the concentration of an inhibitory ligand that it binds (Figure 3–59, red line). Responses of this type are apparently not sharp enough for optimal cell regulation, and most enzymes that are turned on or off by ligand binding consist of symmetric assemblies of identical subunits. With this arrangement, the binding of a molecule of ligand to a single site on one subunit can promote an allosteric change in the entire assembly that helps the neighboring subunits bind the same ligand. As a result, a cooperative allosteric transition occurs (Figure 3–59, blue line), allowing a relatively small change in ligand concentration in the cell to switch the whole assembly from an almost fully active to an almost fully inactive conformation (or vice versa).

The principles involved in a cooperative “all-or-none” transition are the same for all proteins, whether or not they are enzymes. Thus, for example, they are critical for the efficient uptake and release of O2 by hemoglobin in our blood. But they are perhaps easiest to visualize for an enzyme that forms a symmetric dimer. In the example shown in Figure 3–60, the first molecule of an inhibitory ligand binds with great difficulty since its binding disrupts an energetically favorable interaction between the two identical monomers in the dimer. A second molecule of inhibitory ligand now binds more easily, however, because its binding restores the energetically favorable monomer–monomer contacts of a symmetric dimer (this also completely inactivates the enzyme).

As an alternative to this induced fit model for a cooperative allosteric transition, we can view such a symmetric enzyme as having only two possible conformations, corresponding to the “enzyme on” and “enzyme off” structures in Figure 3–60. In this view, ligand binding perturbs an all-or-none equilibrium between these two states, thereby changing the proportion of active molecules. Both models represent true and useful concepts.

Many Changes in Proteins Are Driven by Protein Phosphorylation

Proteins are regulated by more than the reversible binding of other molecules. A second method that eukaryotic cells use extensively to regulate a protein’s function is the covalent addition of a smaller molecule to one or more of its amino acid side chains. The most common such regulatory modification in higher eukaryotes is the addition of a phosphate group. We shall therefore use protein phosphorylation to illustrate some of the general principles involved in the control of protein function through the modification of amino acid side chains.

A phosphorylation event can affect the protein that is modified in three important ways. First, because each phosphate group carries two negative charges, the enzyme-catalyzed addition of a phosphate group to a protein can cause a major conformational change in the protein by, for example, attracting a cluster of positively charged amino acid side chains. This can, in turn, affect the binding of ligands elsewhere on the protein surface, dramatically changing the protein’s activity. When a second enzyme removes the phosphate group, the protein returns to its original conformation and restores its initial activity.

Second, an attached phosphate group can form part of a structure that the binding sites of other proteins recognize. As previously discussed, the SH2 domain binds to a short peptide sequence containing a phosphorylated tyrosine side chain (see Figure 3–40B). More than ten other common domains provide binding sites for attaching their protein to phosphorylated peptides in other protein molecules, each recognizing a phosphorylated amino acid side chain in a different protein context. Third, the addition of a phosphate group can mask a binding site that otherwise holds two proteins together, and thereby disrupt protein–protein interactions. As a result, protein phosphorylation and dephosphorylation very often drive the regulated assembly and disassembly of protein complexes (see, for example, Figure 15–11).

Reversible protein phosphorylation controls the activity, structure, and cellular localization of both enzymes and many other types of proteins in eukaryotic cells. In fact, this regulation is so extensive that more than one-third of the 10,000 or so proteins in a typical mammalian cell are thought to be phosphorylated at any given time—many with more than one phosphate. As might be expected, the addition and removal of phosphate groups from specific proteins often occur in response to signals that specify some change in a cell’s state. For example, the complicated series of events that takes place as a eukaryotic cell divides is largely timed in this way (discussed in Chapter 17), and many of the signals mediating cell–cell interactions are relayed from the plasma membrane to the nucleus by a cascade of protein phosphorylation events (discussed in Chapter 15).

A Eukaryotic Cell Contains a Large Collection of Protein Kinases and Protein

Phosphatases

Protein phosphorylation involves the enzyme-catalyzed transfer of the terminal phosphate group of an ATP molecule to the hydroxyl group on a serine, threonine, or tyrosine side chain of the protein (Figure 3–61). A protein kinase catalyzes this reaction, and the reaction is essentially unidirectional because of the large amount of free energy released when the phosphate–phosphate bond in ATP is broken to produce ADP (discussed in Chapter 2). A protein phosphatase catalyzes the reverse reaction of phosphate removal, or dephosphorylation. Cells contain hundreds of different protein kinases, each responsible for phosphorylating a different protein or set of proteins. There are also many different protein phosphatases; some are highly specific and remove phosphate groups from only one or a few proteins, whereas others act on a broad range of proteins and are targeted to specific substrates by regulatory subunits. The state of phosphorylation of a protein at any moment, and thus its activity, depends on the relative activities of the protein kinases and phosphatases that modify it.

The protein kinases that phosphorylate proteins in eukaryotic cells belong to a very large family of enzymes that share a catalytic (kinase) sequence of about 290 amino acids. The various family members contain different amino acid sequences on either end of the kinase sequence (for example, see Figure 3–10), and often have short amino acid sequences inserted into loops within it. Some of these additional amino acid sequences enable each kinase to recognize the specific set of proteins it phosphorylates, or to bind to structures that localize it in specific regions of the cell. Other parts of the protein regulate the activity of each kinase, so it can be turned on and off in response to different specific signals, as described below.

By comparing the number of amino acid sequence differences between the various members of a protein family, we can construct an “evolutionary tree” that is thought to reflect the pattern of gene duplication and divergence that gave rise to the family. Figure 3–62 shows an evolutionary tree of protein kinases. Kinases with related functions are often located on nearby branches of the tree: the protein kinases involved in cell signaling that phosphorylate tyrosine side chains, for example, are all clustered in the top left corner of the tree. The other kinases shown phosphorylate either a serine or a threonine side chain, and many are organized into clusters that seem to reflect their function—in transmembrane signal transduction, intracellular signal amplification, cell-cycle control, and so on.

As a result of the combined activities of protein kinases and protein phosphatases, the phosphate groups on proteins are continually turning over—being added and then rapidly removed. Such phosphorylation cycles may seem wasteful, but they are important in allowing the phosphorylated proteins to switch rapidly from one state to another: the more rapid the cycle, the faster a population of protein molecules can change its state of phosphorylation in response to a sudden change in its phosphorylation rate (see Figure 15–14). The energy required to drive this phosphorylation cycle is derived from the free energy of ATP hydrolysis, one molecule of which is consumed for each phosphorylation event.

The Regulation of the Src Protein Kinase Reveals How a Protein Can Function as a Microprocessor

The hundreds of different protein kinases in a eukaryotic cell are organized into complex networks of signaling pathways that help to coordinate the cell’s activities, drive the cell cycle, and relay signals into the cell from the cell’s environment. Many of the extracellular signals involved need to be both integrated and amplified by the cell. Individual protein kinases (and other signaling proteins) serve as input–output devices, or “microprocessors,” in the integration process. An important part of the input to these signal-processing proteins comes from the control that is exerted by phosphates added and removed from them by protein kinases and protein phosphatases, respectively.

The Src family of protein kinases (see Figure 3–10) exhibits such behavior. The Src protein (pronounced “sarc” and named for the type of tumor, a sarcoma, that its deregulation can cause) was the first tyrosine kinase to be discovered. It is now known to be part of a subfamily of nine very similar protein kinases, which are found only in multicellular animals.

As indicated by the evolutionary tree in Figure 3–62, sequence comparisons suggest that tyrosine kinases as a group were a relatively late innovation that branched off from the serine/threonine kinases, with the Src subfamily being only one subgroup of the tyrosine kinases created in this way.

The Src protein and its relatives contain a short N-terminal region that becomes covalently linked to a strongly hydrophobic fatty acid, which anchors the kinase at the cytoplasmic face of the plasma membrane. Next along the linear sequence of amino acids come two peptide-binding domains, a Src homology 3 (SH3) domain and an SH2 domain, followed by the kinase catalytic domain (Figure 3–63). These kinases normally exist in an inactive conformation, in which a phosphorylated tyrosine near the C-terminus is bound to the SH2 domain, and the SH3 domain is bound to an internal peptide in a way that distorts the active site of the enzyme and helps to render it inactive.

As shown in Figure 3–64, turning the kinase on involves at least two specific inputs: removal of the C-terminal phosphate and the binding of the SH3 domain by a specific activating protein. In this way, the activation of the Src kinase signals the completion of a particular set of separate upstream events (Figure 3–65). Thus, the Src family of protein kinases serves as specific signal integrators, contributing to the web of information-processing events that enable the cell to compute useful responses to a complex set of different conditions.

Proteins That Bind and Hydrolyze GTP Are Ubiquitous Cell Regulators

We have described how the addition or removal of phosphate groups on a protein can be used by a cell to control the protein’s activity. In the example just discussed, a kinase transfers a phosphate from an ATP molecule to an amino acid side chain of a target protein. Eukaryotic cells also have another way to control protein activity by phosphate addition and removal. In this case, the phosphate is not attached directly to the protein; instead, it is a part of the guanine nucleotide GTP, which binds very tightly to a class of proteins known as GTP-binding proteins. In general, proteins regulated in this way are in their active conformations with GTP bound. The loss of a phosphate group occurs when the bound GTP is hydrolyzed to GDP in a reaction catalyzed by the protein itself, and in its GDP-bound state the protein is inactive. In this way, GTP-binding proteins act as on–off switches whose activity is determined by the presence or absence of an additional phosphate on a bound GDP molecule (Figure 3–66).

GTP-binding proteins (also called GTPases because of the GTP hydrolysis they catalyze) comprise a large family of proteins that all contain variations on the same GTP-binding globular domain. When a tightly bound GTP is hydrolyzed by the GTP-binding protein to GDP, this domain undergoes a conformational change that inactivates the protein. The three-dimensional structure of a prototypical member of this family, the monomeric GTPase called Ras, is shown in Figure 3–67.

The Ras protein has an important role in cell signaling (discussed in Chapter 15). In its GTP-bound form, it is active and stimulates a cascade of protein phosphorylations in the cell. Most of the time, however, the protein is in its inactive, GDP-bound form. It becomes active when it exchanges its GDP for a GTP molecule in response to extracellular signals, such as growth factors, that bind to receptors in the plasma membrane (see Figure 15–47).

Regulatory Proteins GAP and GEF Control the Activity of GTP-Binding Proteins by Determining Whether GTP or GDP Is Bound

GTP-binding proteins are controlled by regulatory proteins that determine whether GTP or GDP is bound, just as phosphorylated proteins are turned on and off by protein kinases and protein phosphatases. Thus, Ras is inactivated by a GTPase-activating protein (GAP), which binds to the Ras protein and induces Ras to hydrolyze its bound GTP molecule to GDP—which remains tightly bound—and inorganic phosphate (Pi), which is rapidly released. The Ras protein stays in its inactive, GDP-bound conformation until it encounters a guanine nucleotide exchange factor (GEF), which binds to GDP-Ras and causes Ras to release its GDP. Because the empty nucleotide-binding site is immediately filled by a GTP molecule (GTP is present in large excess over GDP in cells), the GEF activates Ras by indirectly adding back the phosphate removed by GTP hydrolysis. Thus, in a sense, the roles of GAP and GEF are analogous to those of a protein phosphatase and a protein kinase, respectively (Figure 3–68).

Proteins Can Be Regulated by the Covalent Addition of Other Proteins

Cells contain a special family of small proteins whose members are covalently attached to many other proteins to determine the activity or fate of the second protein. In each case, the carboxyl end of the small protein becomes linked to the amino group of a lysine side chain of a “target” protein through an isopeptide bond. The first such protein discovered, and the most abundantly used, is ubiquitin (Figure 3–69A). Ubiquitin can be covalently attached to target proteins in a variety of ways, each of which has a different meaning for cells.

The major form of ubiquitin addition produces polyubiquitin chains in which—once the first ubiquitin molecule is attached to the target—each subsequent ubiquitin molecule links to Lys48 of the previous ubiquitin, creating a chain of Lys48-linked ubiquitins that are attached to a single lysine side chain of the target protein. This form of polyubiquitin directs the target protein to the interior of a proteasome, where it is digested to small peptides (see Figure 6–84). In other circumstances, only single molecules of ubiquitin are added to proteins. In addition, some target proteins are modified with a different type of polyubiquitin chain. These modifications have different functional consequences for the protein that is targeted (Figure 3–69B).

Related structures are created when a different member of the ubiquitin family, such as SUMO (small ubiquitin-related modifier), is covalently attached to a lysine side chain of target proteins. Not surprisingly, all such modifications are reversible. Cells contain sets of ubiquitylating and deubiquitylating (and sumoylating and desumoylating) enzymes that manipulate these covalent adducts, thereby playing roles analogous to the protein kinases and phosphatases that add and remove phosphates from protein side chains.

An Elaborate Ubiquitin-Conjugating System Is Used to Mark Proteins

How do cells select target proteins for ubiquitin addition? As an initial step, the carboxyl end of ubiquitin needs to be activated. This activation is accomplished when a protein called a ubiquitin-activating enzyme (E1) uses ATP hydrolysis energy to attach ubiquitin to itself through a high-energy covalent bond (a thioester). E1 then passes this activated ubiquitin to one of a set of ubiquitin-conjugating (E2) enzymes, each of which acts in conjunction with a set of accessory (E3) proteins called ubiquitin ligases. There are roughly 30 structurally similar but distinct E2 enzymes in mammals, and hundreds of different E3 proteins that form complexes with specific E2 enzymes.

Figure 3–70 illustrates the process used to mark proteins for proteasomal degradation. [Similar mechanisms are used to attach ubiquitin (and SUMO) to other types of target proteins.] Here, the ubiquitin ligase binds to specific degradation signals, called degrons, in protein substrates, thereby helping E2 to form a polyubiquitin chain linked to a lysine of the substrate protein. This polyubiquitin chain on a target protein will then be recognized by a specific receptor in the proteasome, causing the target protein to be destroyed. Distinct ubiquitin ligases recognize different degradation signals, thereby targeting distinct subsets of intracellular proteins for destruction, often in response to specific signals (see Figure 6–86).

Protein Complexes with Interchangeable Parts Make Efficient Use of Genetic Information

The SCF ubiquitin ligase is a protein complex that binds different “target proteins” at different times in the cell cycle, covalently adding polyubiquitin polypeptide chains to these targets. Its C-shaped structure is formed from five protein subunits, the largest of which serves as a scaffold on which the rest of the complex is built. The structure underlies a remarkable mechanism (Figure 3–71). At one end of the C is an E2 ubiquitin-conjugating enzyme. At the other end is a substrate-binding arm, a subunit known as an F-box protein. These two subunits are separated by a gap of about 5 nm. When this protein complex is activated, the F-box protein binds to a specific site on a target protein, positioning the protein in the gap so that some of its lysine side chains contact the ubiquitin-conjugating enzyme. The enzyme can then catalyze repeated additions of ubiquitin polypeptide to these lysines (see Figure 3–71C), producing polyubiquitin chains that mark the target proteins for rapid destruction in a proteasome.

In this manner, specific proteins are targeted for rapid destruction in response to specific signals, thereby helping to drive the cell cycle (discussed in Chapter 17). The timing of the destruction often involves creating a specific pattern of phosphorylation on the target protein that is required for its recognition by the F-box subunit. It also requires the activation of an SCF ubiquitin ligase that carries the appropriate substrate-binding arm. Many of these arms (the F-box subunits) are interchangeable in the protein complex (see Figure 3–71B), and there are more than 70 human genes that encode them.

As emphasized previously, once a successful protein has evolved, its genetic information tends to be duplicated to produce a family of related proteins. Thus, for example, not only are there many F-box proteins—making possible the recognition of different sets of target proteins—but there is also a family of scaffolds (known as cullins) that give rise to a family of SCF-like ubiquitin ligases.

A protein machine like the SCF ubiquitin ligase, with its interchangeable parts, makes economical use of the genetic information in cells. It also creates opportunities for “rapid” evolution, inasmuch as new functions can evolve for the entire complex simply by producing an alternative version of one of its subunits.

Ubiquitin ligases form a diverse family of protein complexes. Some of these complexes are far larger and more complicated than SCF, but their underlying enzymatic function remains the same (Figure 3–71D).

A GTP-Binding Protein Shows How Large Protein Movements Can Be Generated

Detailed structures obtained for one of the GTP-binding protein family members, the EF-Tu protein, provide a good example of how allosteric changes in protein conformations can produce large movements by amplifying a small, local conformational change. As will be discussed in Chapter 6, EF-Tu is an abundant molecule that serves as an elongation factor (hence the EF) in protein synthesis, loading each aminoacyl-tRNA molecule onto the ribosome. EF-Tu contains a Ras-like domain (see Figure 3–67), and the tRNA molecule forms a tight complex with its GTP-bound form. This tRNA molecule can transfer its amino acid to the growing polypeptide chain only after the GTP bound to EF-Tu is hydrolyzed, dissociating the EF-Tu. Since this GTP hydrolysis is triggered by a proper fit of the tRNA to the mRNA molecule on the ribosome, the EF-Tu serves as a factor that discriminates between correct and incorrect mRNA–tRNA pairings (see Figure 6–65).

By comparing the three-dimensional structure of EF-Tu in its GTP-bound and GDP-bound forms, we can see how the repositioning of the tRNA occurs. The dissociation of the inorganic phosphate group (Pi), which follows the reaction GTP → GDP + Pi, causes a shift of a few tenths of a nanometer at the GTP-binding site, just as it does in the Ras protein. This tiny movement, equivalent to a few times the diameter of a hydrogen atom, causes a conformational change to propagate along a crucial piece of α helix, called the switch helix, in the Ras-like domain of the protein. The switch helix seems to serve as a latch that adheres to a specific site in another domain of the molecule, holding the protein in a “shut” conformation. The conformational change triggered by GTP hydrolysis causes the switch helix to detach, allowing separate domains of the protein to swing apart, through a distance of about 4 nm (Figure 3–72). This releases the bound tRNA molecule, allowing its attached amino acid to be used (Figure 3–73).

Notice in this example how cells have exploited a simple chemical change that occurs on the surface of a small protein domain to create a movement 50 times larger. Dramatic shape changes of this type also cause the very large movements that occur in motor proteins, as we discuss next.

Motor Proteins Produce Large Movements in Cells

We have seen that conformational changes in proteins have a central role in enzyme regulation and cell signaling. We now discuss proteins whose major function is to move other molecules. These motor proteins generate the forces responsible for muscle contraction and the crawling and swimming of cells. Motor proteins also power smaller-scale intracellular movements: they help to move chromosomes to opposite ends of the cell during mitosis (discussed in Chapter 17), to move organelles along molecular tracks within the cell (discussed in Chapter 16), and to move enzymes along a DNA strand during the synthesis of a new DNA molecule (discussed in Chapter 5). All these fundamental processes depend on proteins with moving parts that operate as force-generating machines.

How do these machines work? In other words, how do cells use shape changes in proteins to generate directed movements? If, for example, a protein is required to walk along a narrow thread such as a DNA molecule, it can do this by undergoing a series of conformational changes, such as those shown in Figure 3–74. But with nothing to drive these changes in an orderly sequence, they are perfectly reversible, and the protein can only wander randomly back and forth along the thread. We can look at this situation in another way. Since the directional movement of a protein does work, the laws of thermodynamics (discussed in Chapter 2) demand that such movement use free energy from some other source (otherwise the protein could be used to make a perpetual motion machine). Therefore, without an input of energy, the protein molecule can only wander aimlessly.

How can the cell make such a series of conformational changes unidirectional? To force the entire cycle to proceed in one direction, it is enough to make any one of the changes in shape irreversible. Most proteins that are able to walk in one direction for long distances achieve this motion by coupling one of the conformational changes to the hydrolysis of an ATP molecule that is tightly bound to the protein. The mechanism is similar to the one just discussed that drives allosteric protein shape changes by GTP hydrolysis. Because ATP (or GTP) hydrolysis releases a great deal of free energy, it is very unlikely that the nucleotide-binding protein will undergo the reverse shape change needed for moving backward—since this would require that it also reverse the ATP hydrolysis by adding a phosphate molecule to ADP to form ATP.

In the model shown in Figure 3–75A, ATP binding shifts a motor protein from conformation 1 to conformation 2. The bound ATP is then hydrolyzed to produce ADP and inorganic phosphate (Pi), causing a change from conformation 2 to conformation 3. Finally, the release of the bound ADP and Pi drives the protein back to conformation 1. Because the energy provided by ATP hydrolysis drives the transition 2 → 3, this series of conformational changes is effectively irreversible. Thus, the entire cycle goes in only one direction, causing the protein molecule to walk continuously to the right in this example.

Many motor proteins generate directional movement through the use of a similar unidirectional ratchet, including the muscle motor protein myosin which walks along actin filaments (Figure 3–75B), and the kinesin proteins that walk along microtubules (both discussed in Chapter 16). These movements can be rapid: some of the motor proteins involved in DNA replication (the DNA helicases) propel themselves along a DNA strand at rates as high as 1000 nucleotides per second.

Membrane-Bound Transporters Harness Energy to Pump Molecules Through Membranes

We have thus far seen how proteins that undergo allosteric shape changes can act as microprocessors (Src family kinases), as assembly factors (EF-Tu), and as generators of mechanical force and motion (motor proteins). Allosteric proteins can also harness energy derived from ATP hydrolysis, ion gradients, or electron-transport processes to pump specific ions or small molecules across a membrane. We consider one example here that will be discussed in more detail in Chapter 11.

The ABC transporters (ATP-binding cassette transporters) constitute an important class of membrane-bound pump proteins. In humans, at least 48 different genes encode them. These transporters mostly function to export hydrophobic molecules from the cytoplasm, serving to remove toxic molecules at the mucosal surface of the intestinal tract, for example, or at the blood–brain barrier. The study of ABC transporters is of intense interest in clinical medicine, because the overproduction of proteins in this class contributes to the resistance of tumor cells to chemotherapeutic drugs. In bacteria, the same types of proteins primarily function to import essential nutrients into the cell.

A typical ABC transporter contains a pair of membrane-spanning subunits linked to a pair of ATP-binding subunits located just below the plasma membrane. As in other examples we have discussed, the hydrolysis of the bound ATP molecules drives conformational changes in the protein, transmitting forces that cause the membrane-spanning subunits to move their bound molecules across the lipid bilayer (Figure 3–76).

Humans have invented many different types of mechanical pumps, and it should not be surprising that cells also contain membrane-bound pumps that function in other ways. Among the most notable are the rotary pumps that couple the hydrolysis of ATP to the transport of H+ ions (protons). These pumps resemble miniature turbines, and they are used to acidify the interior of lysosomes and other eukaryotic organelles. Like other ion pumps that create ion gradients, they can function in reverse to catalyze the reaction ADP + Pi → ATP, if the gradient across their membrane of the ion that they transport is steep enough.

One such pump, the ATP synthase, harnesses a gradient of proton concentration produced by electron-transport processes to produce most of the ATP used in the living world. This ubiquitous pump has a central role in energy conversion, and we shall discuss its three-dimensional structure and mechanism in Chapter 14.

Proteins Often Form Large Complexes That Function as Protein Machines

Large proteins formed from many domains are able to perform more elaborate functions than small, single-domain proteins. But large protein assemblies formed from many protein molecules linked together by noncovalent bonds perform the most impressive tasks. Now that it is possible to reconstruct most biological processes in cell-free systems in the laboratory, it is clear that each of the central processes in a cell—such as DNA replication, protein synthesis, vesicle budding, or transmembrane signaling—is catalyzed by a highly coordinated, linked set of 10 or more proteins. In most such protein machines, an energetically favorable reaction such as the hydrolysis of bound nucleoside triphosphates (ATP or GTP) drives an ordered series of conformational changes in one or more of the individual protein subunits, enabling the ensemble of proteins to move coordinately. In this way, each enzyme can be moved directly into position, as the machine catalyzes successive reactions in a series (Figure 3–77). This is what occurs, for example, in protein synthesis on a ribosome (discussed in Chapter 6)—or in DNA replication, where a large multiprotein complex moves rapidly along the DNA (discussed in Chapter 5).

Cells have evolved protein machines for the same reason that humans have invented mechanical and electronic machines. For accomplishing almost any task, manipulations that are spatially and temporally coordinated through linked processes are much more efficient than the use of many separate tools.

Scaffolds Concentrate Sets of Interacting Proteins

As scientists have learned more of the details of cell biology, they have recognized an increasing degree of sophistication in cell chemistry. Thus, not only do we now know that protein machines play a predominant role, but it has also become clear that they are very often localized to specific sites in the cell, being assembled and activated only where and when they are needed. As one example, when extracellular signaling molecules bind to receptor proteins in the plasma membrane, the activated receptors often recruit a set of other proteins to the inside surface of the plasma membrane to form a large protein complex that passes the signal on (discussed in Chapter 15).

The mechanisms frequently involve scaffold proteins. These are proteins with binding sites for multiple other proteins, and they serve both to link together specific sets of interacting proteins and to position them at specific locations inside a cell. At one extreme are rigid scaffolds, such as the cullin in SCF ubiquitin ligase (see Figure 3–71). At the other extreme are the large, flexible scaffold proteins that often underlie regions of specialized plasma membrane. These include the Discs-large protein (Dlg), a protein of about 900 amino acids that is concentrated in special regions beneath the plasma membrane in epithelial cells and at synapses. Dlg contains binding sites for at least seven other proteins, interspersed with regions of more flexible polypeptide chain. An ancient protein, conserved in organisms as diverse as sponges, worms, flies, and humans, Dlg derives its name from the mutant phenotype of the organism in which it was first discovered; the cells in the imaginal discs of a Drosophila embryo with a mutation in the Dlg gene fail to stop proliferating when they should, and they produce unusually large discs whose epithelial cells can form tumors.

Although incompletely studied, Dlg and a large number of similar scaffold proteins are thought to function like the protein that is schematically illustrated in Figure 3–78. By binding a specific set of interacting proteins, these scaffolds can enhance the rate of critical reactions, while also confining them to the particular region of the cell that contains the scaffold. For similar reasons, cells also make extensive use of scaffold RNA molecules, as discussed in Chapter 7.

Many Proteins Are Controlled by Covalent Modifications That Direct Them to Specific Sites Inside the Cell

We have thus far described only a few ways in which proteins are post-translationally modified. A large number of other such modifications also occur, more than 200 distinct types being known. To give a sense of the variety, Table 3–3 presents a few of the modifying groups with known regulatory roles. As in phosphate and ubiquitin additions described previously, these groups are added and then removed from proteins according to the needs of the cell.

A large number of proteins are now known to be modified on more than one amino acid side chain, with different regulatory events producing a different pattern of such modifications. A striking example is the protein p53, which plays a central part in controlling a cell’s response to adverse circumstances (see Figure 17–62). Through one of four different types of molecular additions, this protein can be modified at 20 different sites. Because an enormous number of different combinations of these 20 modifications are possible, the protein’s behavior can in principle be altered in a huge number of ways. Such modifications will often create a site on the modified protein that binds it to a scaffold protein in a specific region of the cell, thereby connecting it—via the scaffold—to the other proteins required for a reaction at that site.

One can view each protein’s set of covalent modifications as a combinatorial regulatory code. Specific modifying groups are added to or removed from a protein in response to signals, and the code then alters protein behavior—changing the activity or stability of the protein, its binding partners, and/or its specific location within the cell (Figure 3–79). As a result, the cell is able to respond rapidly and with great versatility to changes in its condition or environment.

A Complex Network of Protein Interactions Underlies Cell Function

There are many challenges facing cell biologists in this information-rich era when a large number of complete genome sequences are known. One is the need to dissect and reconstruct each one of the thousands of protein machines that exist in an organism such as ourselves. To understand these remarkable protein complexes, each will need to be reconstituted from its purified protein parts, so that we can study its detailed mode of operation under controlled conditions in a test tube, free from all other cell components. This alone is a massive task. But we now know that each of these subcomponents of a cell also interacts with other sets of macromolecules, creating a large network of protein–protein and protein–nucleic acid interactions throughout the cell. To understand the cell, therefore, we will need to analyze most of these other interactions as well.

We can gain some idea of the complexity of intracellular protein networks from a particularly well-studied example described in Chapter 16: the many dozens of proteins that interact with the actin cytoskeleton to control actin filament behavior (see Panel 16–3, p. 905).

The extent of such protein–protein interactions can also be estimated more generally. An enormous amount of valuable information is now freely available in protein databases on the Internet: tens of thousands of three-dimensional protein structures plus tens of millions of protein sequences derived from the nucleotide sequences of genes.

Scientists have been developing new methods for mining this great resource to increase our understanding of cells. In particular, computer-based bioinformatics tools are being combined with robotics and other technologies to allow thousands of proteins to be investigated in a single set of experiments. Proteomics is a term that is often used to describe such research focused on the analysis of large sets of proteins, analogous to the term genomics describing the large-scale analysis of DNA sequences and genes.

A biochemical method based on affinity tagging and mass spectroscopy has proven especially powerful for determining the direct binding interactions between the many different proteins in a cell (discussed in Chapter 8). The results are being tabulated and organized in Internet databases. This allows a cell biologist studying a small set of proteins to readily discover which other proteins in the same cell are likely to bind to, and thus interact with, that set of proteins. When displayed graphically as a protein interaction map, each protein is represented by a box or dot in a two-dimensional network, with a straight line connecting those proteins that have been found to bind to each other.

When hundreds or thousands of proteins are displayed on the same map, the network diagram becomes bewilderingly complicated, serving to illustrate the enormous challenges that face scientists attempting to understand the cell (Figure 3–80). Much more useful are small subsections of these maps, centered on a few proteins of interest.

We have previously described the structure and mode of action of the SCF ubiquitin ligase, using it to illustrate how protein complexes are constructed from interchangeable parts (see Figure 3–71). Figure 3–81 shows a network of protein–protein interactions for the five proteins that form this protein complex in a yeast cell. Four of the subunits of this ligase are located at the bottom right of this figure. The remaining subunit, the F-box protein that serves as its substrate-binding arm, appears as a set of 15 different gene products that bind to adaptor protein 2 (the Skp1 protein). Along the top and left of the figure are sets of additional protein interactions marked with yellow and green shading: as indicated, these protein sets function at the origin of DNA replication, in cell cycle regulation, in methionine synthesis, in the kinetochore, and in vacuolar H+-ATPase assembly. We shall use this figure to explain how such protein interaction maps are used, and what they do and do not mean.

1. Protein interaction maps are useful for identifying the likely function of previously uncharacterized proteins. Examples are the products of the genes that have thus far only been inferred to exist from the yeast genome sequence, which are the three proteins in the figure that lack a simple three-letter abbreviation (white letters beginning with Y). The three in this diagram are F-box proteins that bind to Skp1; these are therefore likely to function as part of the ubiquitin ligase, serving as substrate-binding arms that recognize different target proteins. However, as we discuss next, neither assignment can be considered certain without additional data.

2. Protein interaction networks need to be interpreted with caution because, as a result of evolution making efficient use of each organism’s genetic information, the same protein can be used as part of different protein complexes that have different types of functions. Thus, although protein A binds to protein B and protein B binds to protein C, proteins A and C need not function in the same process. For example, we know from detailed biochemical studies that the functions of Skp1 in the kinetochore and in vacuolar H+-ATPase assembly (yellow shading) are separate from its function in the SCF ubiquitin ligase. In fact, only the remaining three functions of Skp1 illustrated in the diagram—methionine synthesis, cell cycle regulation, and origin of replication (green shading)—involve ubiquitylation.

3. In cross-species comparisons, those proteins displaying similar patterns of interactions in the two protein interaction maps are likely to have the same function in the cell. Thus, as scientists generate more and more highly detailed maps for multiple organisms, the results will become increasingly useful for inferring protein function. These map comparisons will be a particularly powerful tool for deciphering the functions of human proteins, because a vast amount of direct information about protein function can be obtained from genetic engineering, mutational, and genetic analyses in experimental organisms—such as yeast, worms, and flies—that are not feasible in humans.

What does the future hold? There are likely to be on the order of 10,000 different proteins in a typical human cell, each of which interacts with 5 to 10 different partners. Despite the enormous progress made in recent years, we cannot yet claim to understand even the simplest known cells, such as the small Mycoplasma bacterium formed from only about 500 gene products (see Figure 1–10). How then can we hope to understand a human?

Clearly, a great deal of new biochemistry will be essential, in which each protein in a particular interacting set is purified so that its chemistry and interactions can be dissected in a test tube. But in addition, more powerful ways of analyzing networks will be needed based on mathematical and computational tools not yet invented, as we shall emphasize in Chapter 8. Clearly, there are many wonderful challenges that remain for future generations of cell biologists.

Summary

Proteins can form enormously sophisticated chemical devices, whose functions largely depend on the detailed chemical properties of their surfaces. Binding sites for ligands are formed as surface cavities in which precisely positioned amino acid side chains are brought together by protein folding. In this way, normally unreactive amino acid side chains can be activated to make and break covalent bonds. Enzymes are catalytic proteins that greatly speed up reaction rates by binding the high-energy transition states for a specific reaction path; they also can perform acid catalysis and base catalysis simultaneously. The rates of enzyme reactions are often so fast that they are limited only by diffusion. Rates can be further increased only if enzymes that act sequentially on a substrate are joined into a single multienzyme complex, or if the enzymes and their substrates are attached to protein scaffolds, or otherwise confined to the same part of the cell.

Proteins reversibly change their shape when ligands bind to their surface. The allosteric changes in protein conformation produced by one ligand affect the binding of a second ligand, and this linkage between two ligand-binding sites provides a crucial mechanism for regulating cell processes. Metabolic pathways, for example, are controlled by feedback regulation: some small molecules inhibit and other small molecules activate enzymes early in a pathway. Enzymes controlled in this way generally form symmetric assemblies, allowing cooperative conformational changes to create a steep response to changes in the concentrations of the ligands that regulate them.

The expenditure of chemical energy can drive unidirectional changes in protein shape. By coupling allosteric shape changes to ATP hydrolysis, for example, proteins can do useful work, such as generating a mechanical force or moving for long distances in a single direction. The three-dimensional structures of proteins have revealed how a small local change caused by nucleoside triphosphate hydrolysis is amplified to create major changes elsewhere in the protein. By such means, these proteins can serve as input–output devices that transmit information, as assembly factors, as motors, or as membrane-bound pumps.

Highly efficient protein machines are formed by incorporating many different protein molecules into larger assemblies that coordinate the allosteric movements of the individual components. Such machines perform most of the important reactions in cells.

Proteins are subjected to many reversible, post-translational modifications, such as the covalent addition of a phosphate or an acetyl group to a specific amino acid side chain. The addition of these modifying groups is used to regulate the activity of a protein, changing its conformation, its binding to other proteins, and its location inside the cell. A typical protein in a cell will interact with more than five different partners. Through proteomics, biologists can analyze thousands of proteins in one set of experiments. One important result is the production of detailed protein interaction maps, which aim at describing all of the binding interactions between the thousands of distinct proteins in a cell. However, understanding these networks will require new biochemistry, through which small sets of interacting proteins can be purified and their chemistry dissected in detail. In addition, new computational techniques will be required to deal with the enormous complexity.

What we don’t know

- What are the functions of the surprisingly large amount of unfolded polypeptide chain found in proteins?

- How many types of protein functions remain to be discovered? What are the most promising approaches for discovering them?

- When will scientists be able to take any amino acid sequence and accurately predict both that protein’s three-dimensional conformations and its chemical properties? What breakthroughs will be needed to accomplish this important goal?

- Are there ways to reveal the detailed workings of a protein machine that do not require the purification of each of its component parts in large amounts, so that the machine’s functions can be reconstituted and dissected using chemical techniques in a test tube?

- What are the roles of the dozens of different types of covalent modifications of proteins that have been found in addition to those listed in Table 3–3? Which ones are critical for cell function and why?

- Why is amyloid toxic to cells and how does it contribute to neurodegenerative diseases such as Alzheimer’s disease?