TOOLS OF BIOCHEMISTRY 5B

MASS, SEQUENCE, AND AMINO ACID ANALYSES OF PURIFIED PROTEINS

Mass Determination

Once a protein has been purified, how does a researcher convince her/himself that the correct target protein has been obtained in a purified form? The first indication comes typically from an SDS-PAGE gel (see Tools of Biochemistry 2A and Figure 5A.7), which shows (1) the purity of the protein and (2) an approximate molecular weight estimated by comparing the migration of the target protein in the gel to protein molecular weight standards. For example, the molecular weight of the mutant myoglobin shown in Figure 5A.7 is predicted to be 18,232 Da based on the amino acid sequence of the translated gene. Figure 5A.7 shows that the mutant protein migrates between 14 and 22 kDa as would be expected. The purified protein appears to have the correct mass by SDS-PAGE; however, this is not a high resolution technique. The actual mass could differ from the expected mass by several hundred daltons and still appear reasonably close to the expected mass by SDS-PAGE. Mass spectrometry (MS) provides the most accurate mass measurements of large biomolecules. For this reason it is desirable to obtain high resolution mass data via MS to confirm that the protein has no unexpected post-translational modifications (e.g., proteolytic cleavage and/or covalent modifications).

Protein MS has become an indispensable analytical tool since ionization techniques compatible with protein analysis were developed in the late 1980s. The application of MS to significant problems in biochemistry continues to grow as the technology improves. We will present more advanced MS techniques in later chapters; here we will focus on the application of MS to accurate protein mass determination and peptide sequencing.

Figure 5B.1 shows a simplified diagram of a mass spectrometer that contains a single mass analyzer, sufficient for the routine determination of accurate protein masses using **electrospray ionization** (ESI) or **matrix-assisted laser desorption/ionization** (MALDI) techniques. In ESI, a fine mist of protein solution is accelerated toward a mass analyzer. By the time the mist reaches the analyzer most of the solvent has evaporated, leaving protein molecules with a varying number of charges to be separated in the mass analyzer. The detector records the ratio of mass to charge (m/z, where m = mass and z = charge). The ESI-MS mass spectrum is a collection of peaks with different m/z ratios, where m is constant and z varies (Figure 5B.1 top). In the MALDI technique, the protein is imbedded in a large excess (~10,000-fold) of some matrix that absorbs UV light. When a laser pulse hits the matrix, it absorbs the energy of the laser light

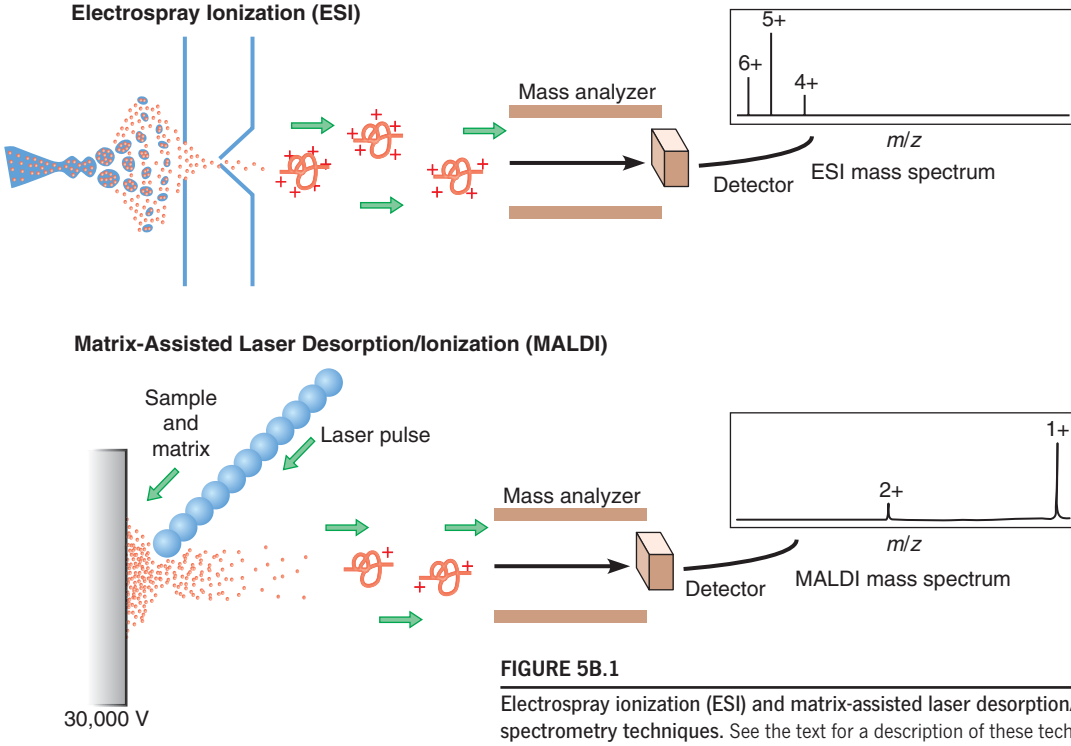


FIGURE 5B.1

Electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) mass spectrometry techniques. See the text for a description of these techniques.

and is vaporized. The vaporized matrix carries intact protein molecules into the gas phase and toward the mass analyzer. The MALDI-MS spectrum shows *m/z* for predominantly the parent ion (Figure 5B.1 bottom).

Sequence Determination

An accurate protein mass is usually sufficient to confirm the identity of a known protein; however, if an unknown protein is the target of some purification scheme, the mass alone is not typically sufficient to identify the protein. In this case, sequence information is also desirable. If the function of the protein is also unknown, the sequence will allow potential identification of the function by similarity searching.

There are several ways in which the amino acid sequence can be determined. As mentioned on page 154, determination of the gene sequence is one of the easiest methods. Indeed, as the entire genomes of many organisms have been determined, we have amino acid sequence information for hundreds of thousands of proteins, many of them of still unknown function. Protein sequences translated from cloned genes do not provide us with information concerning modification of amino acids or the existence of intramolecular cross-links such as disulfide bonds. To find these, we must sequence the protein itself. Here we present two methods for obtaining peptide sequences: tandem MS, which was developed in the mid-1980s and is now the method of choice for most labs, and Edman sequencing, which was developed 20 years earlier by Pehr Edman and is still in use today.

We will present the procedure for Edman sequencing first because there is much useful protein chemistry involved, and the logic for reconstructing the overall protein sequence is the same for both the Edman and tandem MS methods.

The Edman method is based on the stepwise removal of amino acids from the N-terminus of a peptide by a series of chemical reactions called the "Edman degradation" (Figure 5B.2). The compound phenylisothiocyanate (PITC) is reacted in alkali with the N-terminal amino group to yield a phenylthiocarbamyl (PTC) derivative of the peptide (Figure 5B.2, step 1). This derivative is then treated with a strong anhydrous acid, which results in cleavage of the peptide bond between residues 1 and 2 (step 2). The derivative of the N-terminal residue then rearranges to yield a phenylthiohydantoin (PTH) derivative of the amino acid (step 3). Two important things have been accomplished: (1) the N-terminal residue has been marked with an identifiable label, and (2) the rest of the polypeptide has not been destroyed; it has simply been shortened by one residue. The whole sequence of reactions can now be repeated and the second residue determined. By continued repetition, a long polypeptide can be "read," starting from the N-terminal end. This procedure can be performed automatically with an instrument known as a sequenator, which is able to carry out the entire set of reactions shown in Figure 5B.2 over and over again. The sequenator will accumulate in a separate tube the PTH derivative of each amino acid residue in the polypeptide, starting with the N-terminal residue and proceeding for as many cycles as the operator desires or precision allows. The PTH

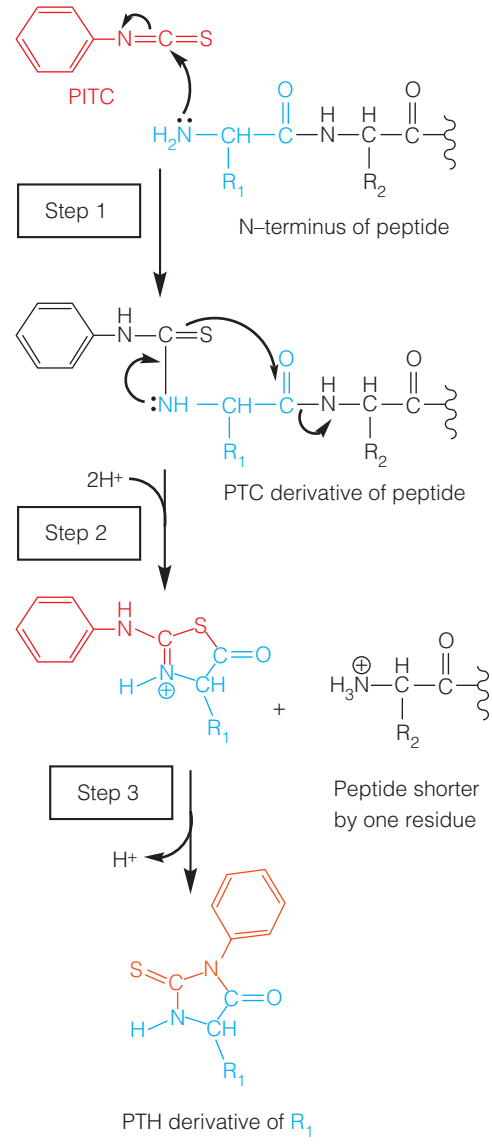


FIGURE 5B.2

The Edman degradation.

derivatives are then identified by high-performance liquid chromatography (HPLC) and/or MS.

In practice, 30–40 residues can be read reliably by Edman sequencing; thus, it is necessary to fragment longer proteins and then sequence the smaller peptides obtained from the fragmentation reaction. Fragmentation is achieved using proteases such as those listed in Table 5.4 and/or cyanogen bromide. Once this first set of fragments has been sequenced, a second fragmentation is carried out using a protease with a different specificity such that a second set of peptides is obtained that overlaps the set obtained in the first fragmentation. We will use the sequencing of bovine insulin as an example to illustrate this process. This choice is appropriate because it was the first protein ever sequenced, by Frederick Sanger and his coworkers in the early 1950s (work for which Sanger won his *first* Nobel Prize). The example is also more complicated than most because we must deal with two covalently

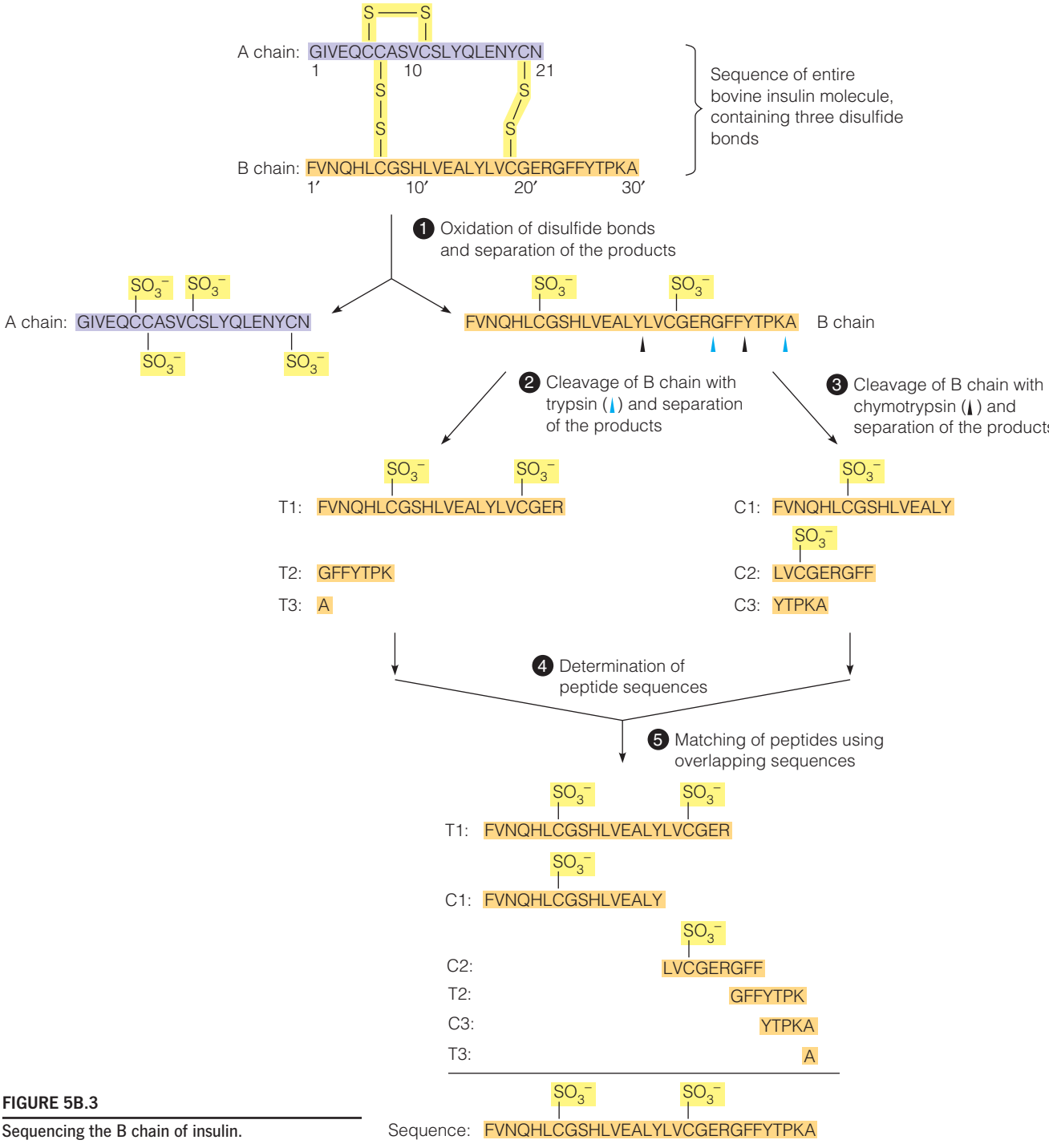
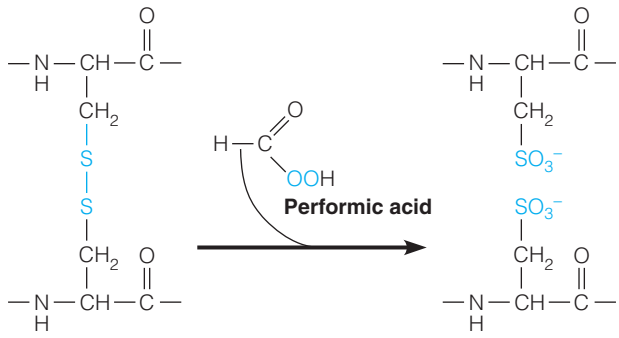


FIGURE 5B.3
Sequencing the B chain of insulin.

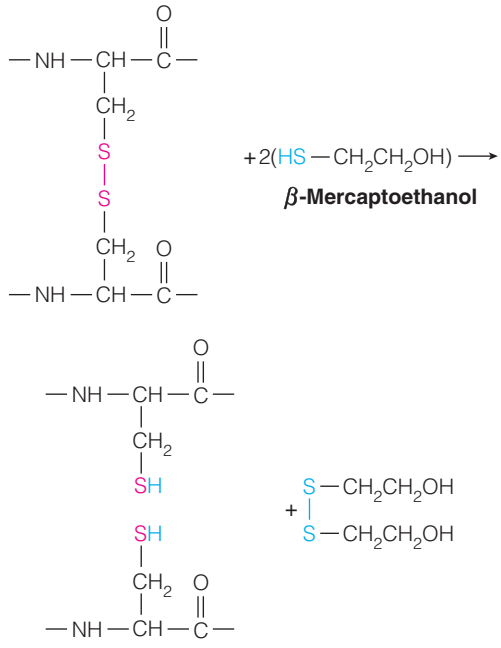
connected chains and locate disulfide bonds. The steps of the procedure are outlined in Figure 5B.3.

The researcher intending to sequence a protein by Edman degradation must first make sure that the material is pure. The protein can be separated from other proteins by some combination of the methods described in Tools of Biochemistry 5A and checked for purity by means of electrophoresis and/or isoelectric focusing. Next, it must be determined whether the material contains more than one polypeptide chain because, in some cases disulfide bridges covalently bond chains together. SDS-PAGE in the presence and absence of reducing agents can answer this question (see Tools of Biochemistry 6B). In the insulin example, there are two chains, A and B, as shown in Figure 5B.3. These chains must be separated and sequenced individually because the Edman degradation would release two sets of PTH derivatives simultaneously if the peptides were not separated. To break disulfide bonds and thus separate the chains, several reactions are available. Descriptions of two common procedures follow.

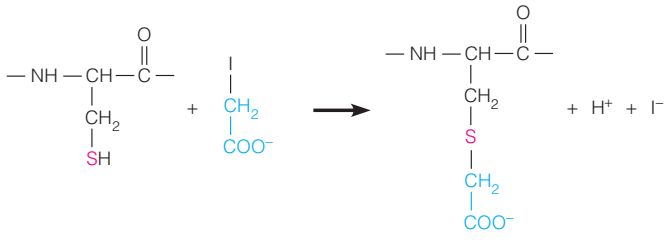
Performic acid oxidation is the technique used in Figure 5B.3, step 1. The strong oxidizing agent performic acid will *irreversibly* react with cystine to yield cysteic acid residues:



Reduction with β-mercaptoethanol is a milder, and *reversible*, technique.



Reduction leaves free sulfhydryl groups, often positioned so that reoxidation to re-form the disulfide bond is likely. Therefore, the sulfhydryls are usually blocked to prevent this. A common blocking reagent is iodoacetate:



If either of these methods is carried out with insulin, the intact protein is cleaved into A and B chains. These chains can then be separated by chromatographic methods.

Before Edman sequencing of the individual chains is started, their amino acid composition is usually determined (see below). This determination may point to unusual compositions and thus warn the operator of potential problems. Furthermore, composition data will serve as a check on the sequencing results because the sequence determined must be consistent with the amino acid composition.

In bovine insulin, the A and B chains are so short that modern instrumentation could sequence either in one sequenator run; however, to demonstrate the methods needed for larger proteins, we assume that the investigator must cleave the insulin chains into shorter polypeptides (this was indeed the case in Sanger's pioneering sequencing studies on insulin). Suppose the insulin B chain is to be sequenced. A first step would be to cleave separate aliquots of the chain with two or more of the specific cleavage reagents described in Table 5.4. Trypsin and chymotrypsin, for instance, would yield the sets of peptides shown in steps 2 and 3, respectively, of Figure 5B.3. The individual peptides would then be isolated from each of the two mixtures, using, for example, ion-exchange chromatography, and their sequences could be determined (Figure 5B.3, step 4).

Suppose each of the peptides shown in Figure 5B.3 has been sequenced (step 4). Although the tryptic peptides alone cover the whole sequence, they are not sufficient to allow us to write down the sequence of the insulin B chain because we do not know the order in which they appear in the intact chain. To overcome this problem we also have the chymotryptic peptides, which overlap the tryptic peptides; therefore, all ambiguity is removed. Only one arrangement of the whole chain is consistent with the sequences of these two sets of peptides, as can be seen by matching overlapping sequences (step 5).

Finally, a complete characterization of the covalent structure of a protein requires that the positions of any disulfide bonds be located. In preparation for sequencing, these bonds would have been destroyed, but the positions of all cysteines, some of which *might* have been involved in bonding, would have been determined. How can we determine which cysteines are linked via disulfide bonds in the native protein?

To determine the arrangement of disulfide bonds, the experimenter again starts with the native protein—insulin in the example shown in Figure 5B.4. Reaction with radioactively labeled iodoacetate marks any free cysteine residues, and fragmentation of the protein into the same peptides used in sequencing allows the positions of these nonbonded cysteines to be identified (step 1). Then samples of the intact protein are cut with various cleavage reagents but now without first cleaving disulfide bonds (steps 2 and 3). Some peptides, which are connected by these bonds, are attached to one another. These can then be isolated and their disulfide bonds cleaved to map the location of each disulfide-bonded cysteine in the protein.

Mass spectrometry can also be used to obtain peptide-sequence information. To do this the mass spectrometer must have a collision cell and two mass analyzers rather than the single mass analyzer shown in Figure 5B.1. Figure 5B.5 shows a schematic diagram of a tandem mass spectrometer (MS-MS) capable of peptide sequencing, where the two mass analyzers are labeled "quadrupole analyzer" and "time-of-flight analyzer." The role of each mass analyzer will be described below.

As in Edman sequencing, MS-MS sequencing works best on smaller peptide fragments. The fragments can be generated

FIGURE 5B.4

Locating the disulfide bonds in insulin.

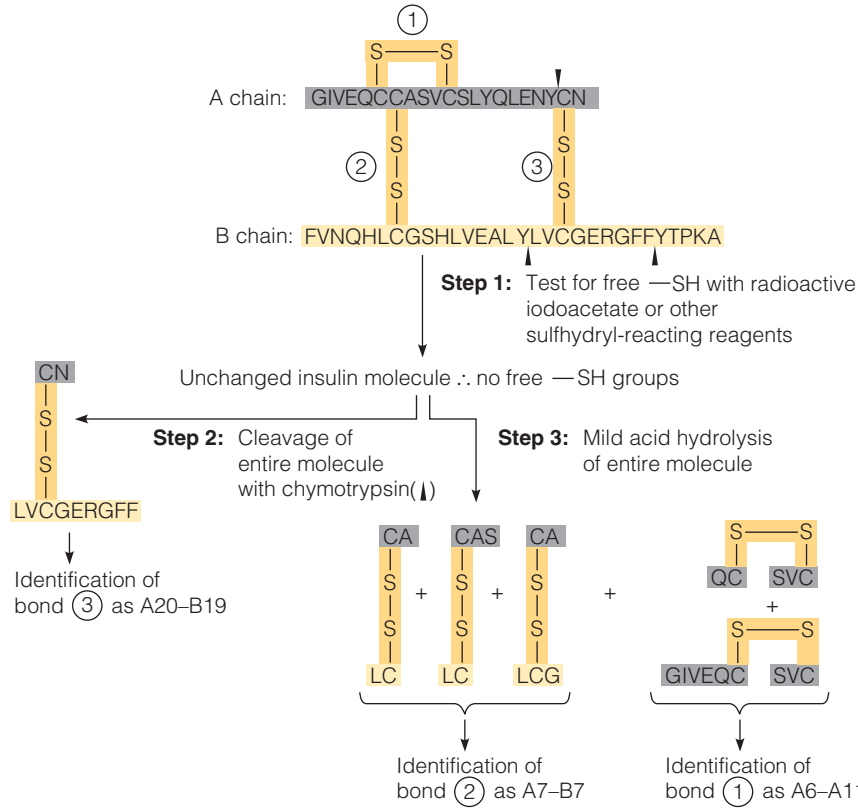
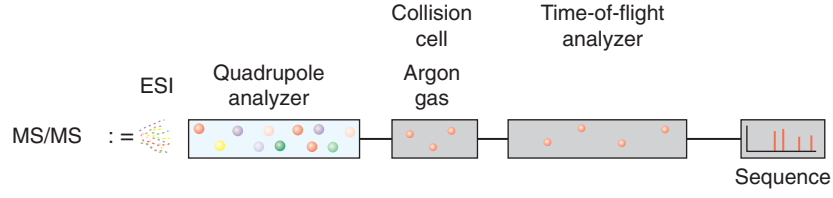


FIGURE 5B.5

Peptide sequencing by MS-MS.



using proteases, or in the spectrometer itself. Let us consider the case in which the fragments are generated using a protease and then introduced into the MS-MS instrument using electrospray. Electrospray is advantageous for sequencing because the fragments tend to have multiple charges distributed along their lengths. Recall that the mass detector records *m/z*; thus, without a charge the fragment is not detectable. As depicted in Figure 5B.5 the mixture of peptide fragments is introduced by electrospray into the first mass analyzer (the quadrupole analyzer). The quadrupole analyzer can be tuned to select a specific fragment (i.e., a specific *m/z* range) for introduction into the collision cell. In the course of the analysis each fragment in turn will be directed from the quadrupole analyzer into the collision cell.

In the collision cell the selected fragment is fragmented further by collisions with argon atoms. To a large extent the fragmentations in the collision cell result in cleavage of the peptide backbone as shown in Figure 5B.6, where cleavage of the first peptide bond gives two subfragments: the N-terminal subfragment that includes the residue R_1, and the C-terminal subfragment that includes residues R_2–R_4. By convention, the N-terminal subfragments are called "*b* ions" and the C-terminal subfragments are called "*y* ions."

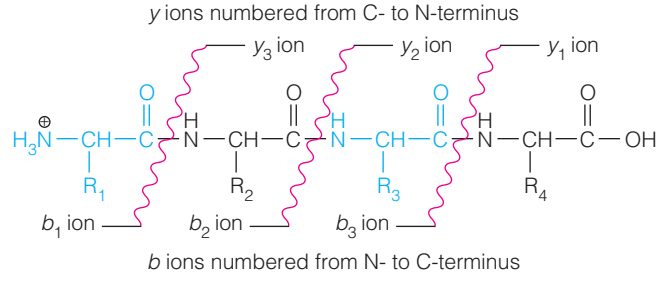


FIGURE 5B.6

Principal ions generated by low energy collision-induced dissocation. The wavy red lines indicate sites of peptide bond cleavage in the collision cell (see Figure 5B.5).

In the collision cell two series of subfragments are generated simultaneously—a series of *b* ions and the corresponding set of *y* ions. Within each series the masses of the ions differ from one another by the mass of a single amino acid residue (i.e., in Figure 5B.6 the masses of b_2 and b_3 differ by the mass of residue R_3). The *m/z* ratio for each ion is determined in the time-of-flight analyzer and recorded to generate a complex spectrum with peaks for each

ion. Because the masses for amino acid residues are known (see Table 5.1), the amino acids present in each fragment can be reliably identified. Modern MS-MS instruments include software that can rapidly identify fragmentation patterns consistent with a specific amino acid sequence. This process is repeated until every fragment that enters the quadrupole analyzer is sequenced (a matter of a few minutes), thereby generating a set of peptide sequences for the protein. To find the order of the peptides a second MS-MS analysis is performed on a series of different fragments generated by a different protease (or CNBr, etc.).

Both MS-MS and Edman sequencing are used to determine peptide sequences and each has its set of strengths and weaknesses. Edman sequencing requires a few micrograms $(10^{-6}\,g)$ of purified protein. It gives nearly complete sequence coverage; however, if the N-terminus of the intact protein is modified, the Edman degradation is blocked. This problem is somewhat ameliorated by fragmentation methods that generate unmodified N-termini for each fragment. MS-MS methods are very sensitive; thus, only picograms $(10^{-12}\,g)$ of protein are needed. Also, separation of chains or fragments is not required because the MS-MS method achieves fragment separation in the mass spectrometer; thus, the MS-MS method is rapid. MS-MS does not usually give complete sequence coverage because some amino acid sequences are difficult to ionize; however, 70–80% sequence coverage is typical (recall that only ions can be detected in the MS instrument—if a peptide carries no charge it will not be detected).

We have described how the entire amino acid sequence, or primary structure, of a protein can be determined. Such analyses have been carried out on several thousand different proteins in the years since Sanger first determined the sequence of insulin. Today it is rare that an entire protein sequence would be determined using these methods because the sequencing of genes is much more rapid (and frequently precedes the isolation of the protein of interest); however, MS-MS is often used to determine a portion of the sequence of an unknown protein. With a sequence of only 6–10 amino acids one can often identify a protein by searching databases of protein sequences. This use of protein sequence information is the basis for the field of **proteomics**, which we discuss briefly in Tools of Biochemistry 5D.

Amino Acid Analysis

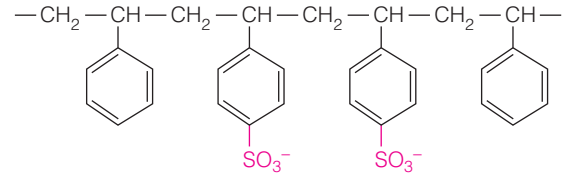
Finally, we turn our attention to the determination of the amino acid composition (or "amino acid analysis") of a purified protein. Given recent developments in gene sequencing and mass spectrometry of proteins, the determination of the amino acid composition of a protein is no longer a common analytical procedure; however, it remains a standard method for the accurate quantitation of protein in an analytical sample.

Amino acid analysis (AAA) involves three basic steps:

1. *Hydrolysis* of the protein to its constituent amino acids.
2. *Separation* of the amino acids in the mixture.
3. *Quantitation* of the individual amino acids.

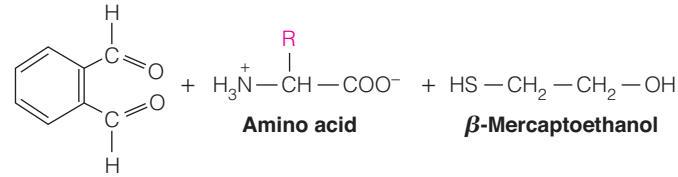
A small sample of the protein is first purified, perhaps by some combination of the methods described in Tools of Biochemistry 5A. The purified protein is dissolved in 6 M HCl, and the solution is sealed in an evacuated ampoule. It is then heated at 105–110 °C for about 24 hours. Under these conditions, the metastable peptide bonds between the residues are completely hydrolyzed.

The hydrolyzed sample is then separated into the constituent amino acids on a cation-exchange column. The kinds of resin typically used are sulfonated polystyrenes:



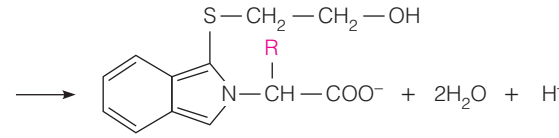
Such a resin separates amino acids in two ways. First, because it is negatively charged, it tends to pass acidic amino acids first and retain basic ones. The pH of the eluting buffer is increased during elution to facilitate this separation. Second, the hydrophobic nature of the polystyrene itself tends to hold up the more hydrophobic amino acids such as leucine and phenylalanine. An example of such an analysis is shown in Figure 5B.7. Note the order of appearance of the amino acids, proceeding from the more acidic to the more basic. Modern amino acid analyzers are completely automated and carry out both the chromatographic separation of the amino acids and their quantitation.

There are many methods for detection and quantitation of the amino acids eluting from the column; but, fluorescence is commonly used due to its sensitivity. For example, the amino acids may be reacted with *o*-phthalaldehyde to yield a fluorescent complex:



Amino acid **β-Mercaptoethanol**

***o*-Phthalaldehyde**



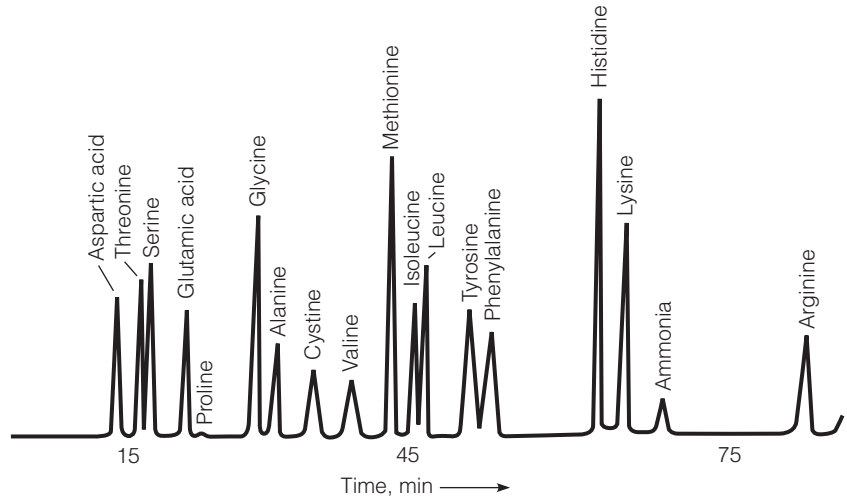
Isoindole derivative of amino acid

Such detection techniques easily give sensitivity to the picomole (pmol, or 10^{-12} mol) range. Microelectrophoresis systems and fluorescence detection have extended this sensitivity to the *attomole* (amol, or 10^{-18} mol) range. This amount corresponds to only a few thousand molecules. Indeed, amino acid analysis techniques have proceeded to the point that the amount of protein contained in one spot in two-dimensional gel electrophoresis (see Figure 1.11, page 19) can be analyzed easily.

Of course, these procedures are not as simple and trouble free as the foregoing discussion might imply. Some amino acids give problems in reaction with the compounds used for detection; proline in particular because it is a secondary amino acid, often reacts

FIGURE 5B.7

Analysis of a protein hydrolysate on a single-column amino acid analyzer. The chromatogram shows the order of elution of hydrolyzed amino acids on a polystyrene column. Free amino acids are detected by absorbance at 220 nm.



slowly or not at all. Furthermore, some amino acids tend to be partially destroyed during the severe hydrolysis. Tryptophan is troublesome in this respect and must be determined by hydrolysis with base and detection by ultraviolet absorbance (see Figure 5.6). Serine, threonine, and tyrosine also tend to be degraded during long hydrolysis. To a considerable extent, these difficulties can be circumvented either by carrying out protective reactions first or by measuring the apparent content of the amino acid at different hydrolysis times and extrapolating to zero hydrolysis time. Asparagine and glutamine are invariably hydrolyzed to aspartic and glutamic acids, so that the total content of these acids observed includes the amides. This reaction, as well as the other degradation reactions mentioned above, can be avoided by using an enzymatic hydrolysis, with a mixture of proteolytic enzymes, in place of the acid hydrolysis. However, this method also has its drawbacks because it is sometimes difficult to achieve complete hydrolysis and the enzymes themselves must be removed before analysis. Despite such complications, amino acid analysis, using automated analyzers, has become a routine operation in protein characterization.

References

Cañas, B., D. López-Ferrer, A. Ramos-Fernández, E. Camafeita, and E. Calvo (2006) Mass spectrometry technologies for proteomics. *Brief. Funct. Genom. Proteom.* 4:295–320.

Cheng, Y.-F., and N. Dovichi (1988) Subattomole amino acid analysis by capillary zone electrophoresis and laser-induced fluorescence. *Science* 242:562–564.

Edman, P., and G. Begg (1967) A protein sequenator. *Eur. J. Biochem.* 1:80–91. The first automated method.

Liu, T.-Y. (1972) Determination of tryptophan. *Methods Enzymol.* 25:44–55.

Thomas, J. J., R. Bakhtiar, and G. Suizdak (2000) Mass Spectrometry in Viral Proteomics. *Acc. Chem. Res.* 33:179–187.

Walsh, K. A., Ericsson, L. H., Parmelee, D. C., and K. Titani (1981) Advances in protein sequencing. *Annu. Rev. Biochem.* 50:261–284.

See also this Website, maintained by A. E. Ashcroft, describing mass spectrometry: www.astbury.leeds.ac.uk/facil/MStut/mstutorial.htm

TOOLS OF BIOCHEMISTRY 5C

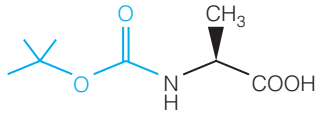
HOW TO SYNTHESIZE A POLYPEPTIDE

Chemical synthesis of peptides of defined sequence is of great importance in medicine and molecular biology. Some synthetic hormones can be made with non-natural amino acids that make them more stable in vivo, and therefore better therapeutics. Synthetic peptides can be used to elicit antibodies against portions of specific proteins; such antibodies are useful in studying the interaction of proteins with other molecules.

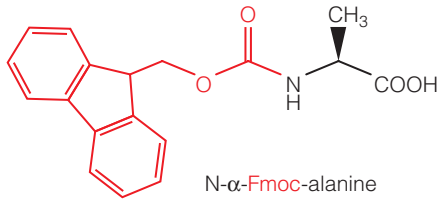
To synthesize a peptide of defined sequence, several criteria must be met:

1. It should be possible to add amino acids one at a time, preferably in an automated reactor.
2. Because of peptide-bond metastability, the amino acids must be activated in some way such that peptide-bond formation is both favorable and efficient (>98% per cycle).
3. To avoid side reactions, all reactive groups (in this case good nucleophiles) must be protected (i.e., blocked from reacting) other than the carboxyl and amino groups that are meant to form the desired peptide bond.
4. The protecting groups used for side chain groups must be stable for the entire synthesis; but, the protecting group for the α-amino group must be removed selectively for each cycle of peptide bond formation.

There are two common synthetic schemes used to make peptides. The solid-phase peptide synthesis (SPPS) chemistry developed by Bruce Merrifield (and recognized by the Nobel Prize) uses the *t*-butyloxycarbonyl (Boc) group to protect the α-amino group. A second popular scheme uses the 9-fluorenylmethoxylcarbonyl (Fmoc) group to protect the α-amino group. Both schemes are amenable to automated solid-phase synthesis and are still widely used. In some cases, one scheme will give a better yield for a given sequence than another, so most labs that synthesize peptides have machines for both Boc and Fmoc methods.



N-α-*t*Boc-alanine



N-α-Fmoc-alanine

A simplified scheme based on the Merrifield solid-phase chemistry is shown in Figure 5C.1. The advantage of the solid-phase method is that the growing peptide chain remains attached to a solid matrix (called the "SPPS resin") until the last step of the synthesis; thus, during each step, excess reactants and contaminants can be washed away. To begin the synthesis, the C-terminal amino acid of the desired peptide sequence is covalently attached to a bead of SPPS resin, with its α-amino group exposed. A series of three steps is then carried out to complete a cycle of peptide-bond formation. These steps include: (1) deprotonation of the α-amino group to make it a better nucleophile; (2) activation of the carboxylic acid group of the next amino acid in the desired sequence, followed by covalent addition of the activated amino acid to the growing peptide chain; and (3) deprotection of the new N-terminal α-amino group. The new peptide bond is formed in step 2. Steps 1–3 are repeated until the desired sequence has been synthesized. Finally, the N-terminal α-amino group and all the protected side chains are deprotected, and the peptide is cleaved from the resin.

A brief description of each step in a peptide-bond synthesis cycle follows. The amino acid to be added to the growing peptide bound to the resin has a free carboxylic acid group, a Boc-protected α-amino group and, if necessary, a protected side chain group. The carboxylic acid is converted in situ to a more reactive carboxylic ester by reaction with a carbodiimide reagent (step 2). The coupling reaction yields a new peptide bond and a peptide, longer by one amino acid, that has a Boc-protected α-amino group. The Boc protecting group is removed selectively using trifluoroacetic acid (step 3), and the next activated residue is then added (repeats steps 1–3). Note that the last residue added is the N-terminal residue. All reactions are carried out automatically, with the growing chains attached to the resin. In the final step, HF is added to remove any side-chain protecting groups and simultaneously cleave the peptide from the resin.

Using these methods, peptides of 50 residues in length can be routinely synthesized in good yield. Expert laboratories can make chains of roughly 150 amino acids. Merrifield, for example, synthesized an active enzyme (ribonuclease) of 124 residues, and Stephen Kent and coworkers have made small proteins of 140–160 amino acids. There are also methods for condensing a synthetic peptide with a sequence derived from an intact biosynthetic protein. In this way non-natural amino acids can be incorporated into larger proteins. A description of these methods, called "native chemical ligation" and "expressed protein ligation," is beyond the scope of this discussion. The interested reader can find more details in the citations given at the end of this section.

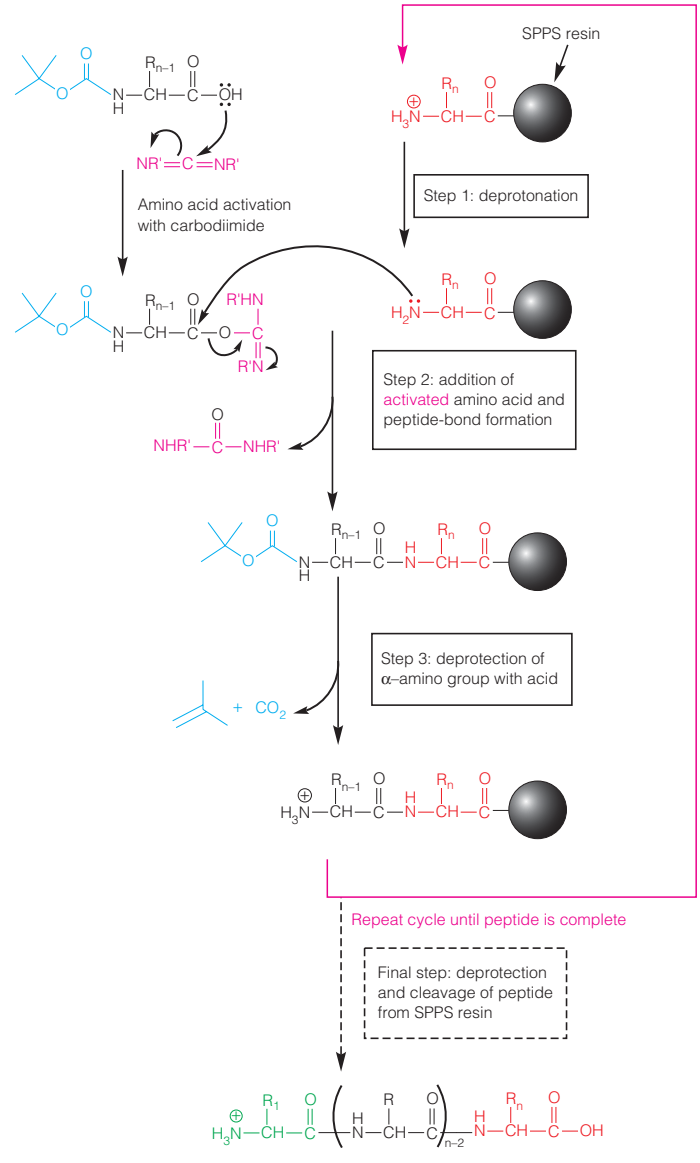


FIGURE 5C.1

Reactions in solid-phase peptide synthesis. Individual steps are described in the text.

Spatially Organized Combinatorial Peptide Arrays

Frequently, it is necessary to test simultaneously a large number of different peptides for some kind of biological activity. One might want to know, for example, which member(s) of a large family of similar oligopeptides is(are) the antigen(s) reacting with a specific antibody. Formerly, this was an extremely laborious process, involving perhaps hundreds of separate syntheses.

Using techniques borrowed from photolithography and inkjet printing, it is now possible to prepare microscopic, two-dimensional arrays containing many combinations of peptides grown on a solid surface. The photolithographic technique is illustrated in Figure 5C.2. The amino acids to be used are each blocked on the N-terminus with a photolabile protecting group and carry activated carboxyl groups. First, one class of amino acids (in this case Leu) is reacted with a surface coated with amino groups. The whole surface is then illuminated, which removes the protective groups. A second activated amino acid can then be added to each chain. In this example, after four rounds, the peptide GGFL has been grown on each site. To generate sequence diversity, a rectangular mask is placed over the surface, so that only half the squares in a checkerboard pattern are illuminated. This allows coupling Tyr residues in the illuminated portion. The other portion is then illuminated and coupled with Pro. Thus, in this example, a simple checkerboard pattern is obtained, with PGGFL and YGGFL alternating. Figure 5C.3 shows the reaction of a fluorescent antibody reactive to YGGFL on such a surface. The example shown is simple: much more complex patterns can easily be generated by use of overlapping masks, allowing thousands of different peptides to be generated, in a prescribed pattern, on one surface.

Using a different synthetic strategy, protein microarrays with greater than 10,000 different proteins immobilized on a single glass slide have been developed for the rapid detection of protein–protein interactions and the determination of intracellular protein expression levels. These "protein chips" are created by depositing a few nanoliters of a protein solution at a precise location on the slide. Each spot on such a slide can be a different protein. The proteins in these microarrays are capable of interacting with other proteins and smaller molecules such as drug candidates and enzyme substrates; thus, this technology allows for direct measurement of protein–protein and protein–ligand interactions.

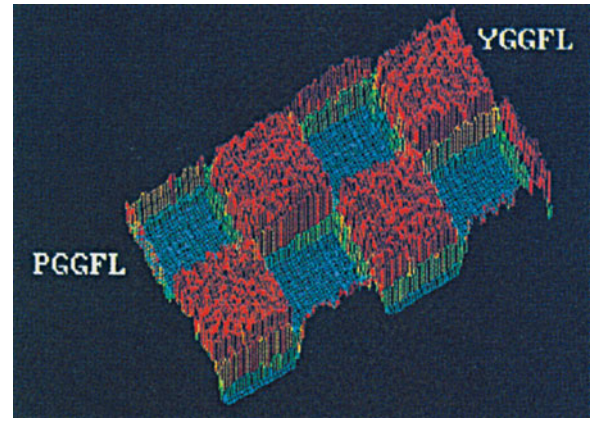


FIGURE 5C.3

Three-dimensional representation of the checkerboard array of YGGFL and PGGFL. Fluorescence intensity data were converted into spike heights that are proportional to the number of counts detected from 2.5-mm square pixels. The spikes are also color coded.

From *Science* 251:767–773, S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, Light-directed, spatially addressable parallel chemical synthesis. © 1991. Reprinted with permission from AAAS and Stephen P. A. Fodor.

References

Clark-Lewis, I., R. Aebersold, H. Ziltener, J. W. Schrader, L. E. Hood, and S. B. H. Kent (1986) Automated chemical synthesis of a protein growth factor for hemopoietic cells, interleukin-3. *Science* 231:134–139.

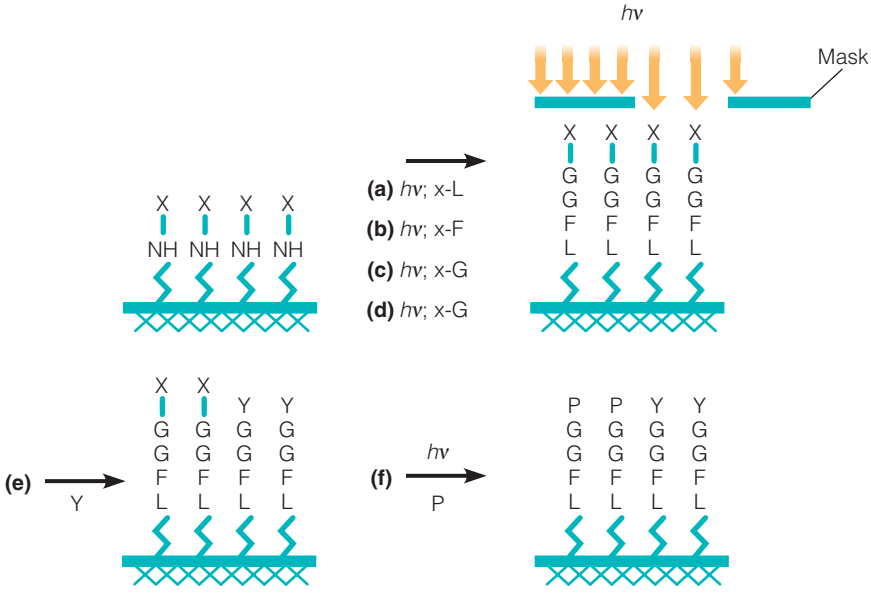


FIGURE 5C.2

An example of light-directed, spatially patterned oligopeptide synthesis. X indicates a photolabile blocking group attached to each amino acid residue added. In steps a–d, the tetrapeptide GGFL is built up on the surface. A mask is then used to illuminate and cleave blocking groups in defined areas, to allow addition of tyrosine (step e). An additional round of photo deprotection and coupling is used to add proline in the remaining areas.

From *Science* 251:767–773, S. P. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas, Light-directed, spatially addressable parallel chemical synthesis. © 1991. Reprinted with permission from AAAS and Stephen P. A. Fodor.

Dawson, P. E., T. W. Muir, I. Clark-Lewis, and S. B. H. Kent (1994) Synthesis of proteins by native chemical ligation. *Science* 266:776–779.
Flavell, R. R., and T. W. Muir (2009) Expressed protein ligation (EPL) in the study of signal transduction, ion conduction, and chromatin biology. *Accts Chem. Res.*42:107–116.
Fodor, S. P. A., Reed, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and D. Solas (1991) Light-directed spatially addressable parallel chemical synthesis. *Science* 251:767–773.
Kochendoerfer, G. G., et al. (2003) Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. *Science* 299:884–887.
MacBeath, G., and S. L. Schreiber (2000) Printing proteins as microarrays for high-throughput function determination. *Science* 289:1760–1763.
Merrifield, B. (1986) Solid phase synthesis. *Science* 232:341–347.
Schnolzer, M., and S. B. Kent (1992) Constructing proteins by dovetailing unprotected synthetic peptides: Backbone-engineered HIV protease. *Science* 256:221–225.
Vila-Perelló, M., and T. W. Muir (2010) Biological applications of protein splicing. *Cell* 143:191–200.
Zhu, H., M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder (2001) Global analysis of protein activities using proteome chips. *Science* 293:2101–2105.

TOOLS OF BIOCHEMISTRY 5D

A BRIEF INTRODUCTION TO PROTEOMICS

The complement of proteins present in a given cell make up the so-called **proteome** of that cell. As mentioned in Chapter 1, **proteomics** is the field of study that attempts to understand the complex relationships between proteins and cell function through global analysis of the proteome rather than investigating the properties of purified protein in isolation. Proteomics includes, among other things, efforts to understand how protein expression and/or post-translational modification levels change in cells, and the consequences of such changes. For example, how does the proteome of a normal pancreatic cell differ from the proteome of a cancerous pancreatic cell? How might a researcher begin to address this question since every cell has the potential to express many thousands of different proteins at any given time? The protein chips described in Tools of Biochemistry 5C offer one method for obtaining such information; but, there are other methods which are more commonly used.

A proteomics experiment can, for example, rapidly identify differences in the levels of cellular expression, or post-translational modification (e.g., phosphorylation) between the proteomes of normal and diseased cells. The key is to correctly identify the affected proteins. The best technique for achieving this is mass spectrometry (see Tools of Biochemistry 5B). Once the affected proteins have been identified, the putative role of the protein in the disease can be researched.

A typical proteomics experiment includes the following steps: (1) separation and isolation of proteins, or protein fragments, from cells or an organism; (2) identification by MS-MS sequencing (Figures 5B.5 and 5B.6) of a particular protein within the complex mixture; and (3) database searching to identify the target protein, and its putative function. In practice, 2-D electrophoresis can be used to separate peptides (see Figure 1.11, page 19) as well as to identify potential protein targets for proteomic analysis. However, the extraction of peptides from 2-D gels is laborious, and direct analysis of complex peptide mixtures using tandem mass spectrometry is preferred for large-scale proteomic analyses.

A basic proteomics experiment is illustrated in Figure 5D.1. A complex mixture of proteins is digested—in this case with the protease trypsin. This mixture gives rise to a complex mixture of peptides that can be separated, for example, by HPLC (see Tools of Biochemistry 5B). Here, the HPLC effluent is injected directly into a mass spectrometer. The complexity of this peptide mixture is represented by the many peaks in the total ion current (TIC) mass spectrum [Panel (a)]. If a very complex sample (e.g., a cellular extract) is the starting material for this experiment, it is likely that each of the peaks in the TIC mass spectrum will contain several peptides; however, these peptides can be separated within the mass spectrometer based on differences in mass-to-charge ratio between the peptides. The separated peptides are shown in a much less complex "parent ion" mass spectrum [Panel (b)]. Each parent ion can be analyzed by MS-MS to yield amino acid sequence (see Figures 5B.5 and 5B.6). Panel (c) of Figure 5D.1 shows the MS-MS spectrum of one of the parent ions from the mass spectrum in Panel (b). The experimentally determined amino acid sequence is then used as input to search a protein sequence database. In the example here, the sequence of the peptide is a fragment of bovine serum albumin.

MS-MS is compatible with the analysis of complex peptide mixtures because only one peptide fragment from the mixture is selected for sequencing by the mass analyzer. Thus, it is still possible to make a positive identification of a protein from complex mixtures of many proteins. In theory, all the proteins in the mixture can be identified, assuming the sequences are listed in some database.

Mass spectrometry is particularly suited to the detection of post-translational modifications. A common modification is protein phosphorylation, which confers changes in both mass and charge on the phosphorylated protein. The presence of some enzymatic activity can also be detected by mass analysis by

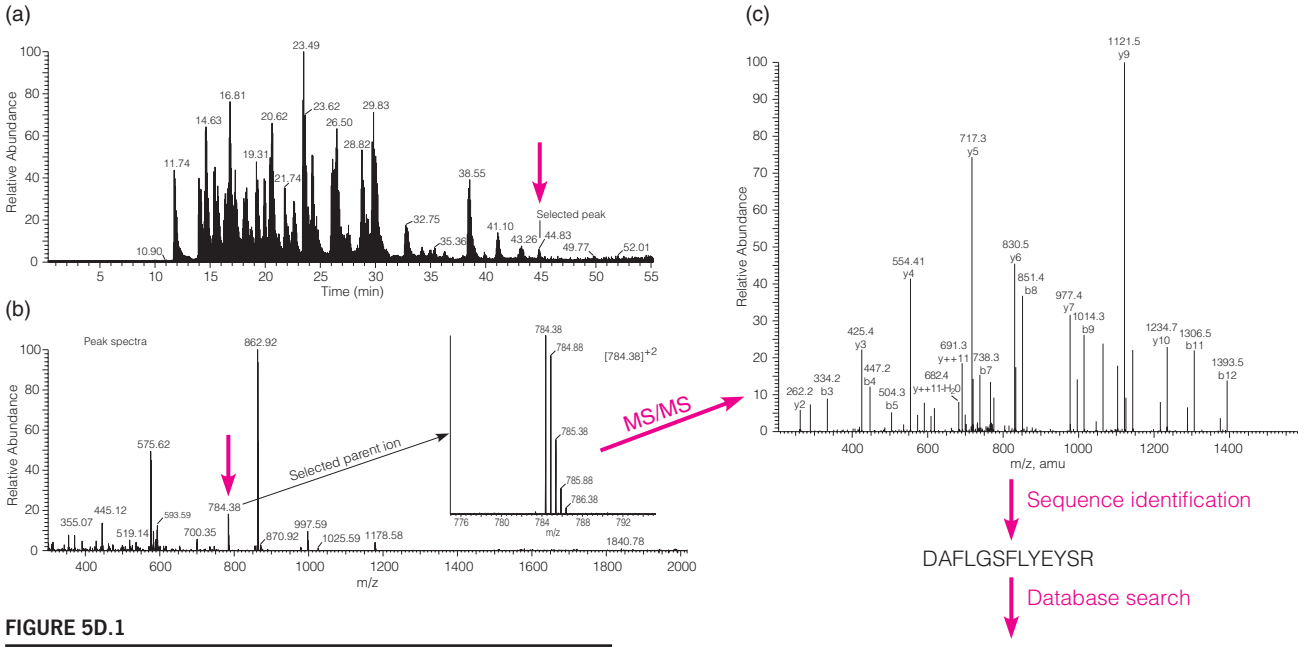


FIGURE 5D.1

Identification of a protein of interest using proteomics methods. Panel **(a)** A total ion current (TIC) mass spectrum for a complex mixture of peptides separated by HPLC. The peptide ions from any portion of the TIC mass spectrum can be selected within the mass spectrometer and subjected to further analysis by tandem mass spectroscopy (e.g., the peak under the red arrow). Panel **(b)** Fragmentation of the ions selected in Panel (a) gives a set of ions that can be further separated by mass-to-charge ratio in the mass spectrometer. Panel **(c)**: An MS-MS spectrum for one of the parent ions in Panel (b) [red arrow in Panel (b)]. The 13 amino acid sequence obtained from MS-MS analysis is used in a database search.

Panels (a–c) courtesy of Jack Benner. Panel (d) courtesy of SIB Swiss Institute of Bioinformatics.

P02769 Serum albumin precursor (Allergen Bos d 6) (BSA)

MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGL
VLIAFSQYLQQCPFDEHVKLVNELTEFAKTCVADESHAGCEKSLHT
LFGDELCKVASLRETYGDMADCCEKQEPERNECFLSHKDDSPDL
PKLKPDPNTLCDEFKADEKKFWGKYLYEIARRHPYFYAPELLYYAN
KYNGVFQECCQAEDKGACLLPKIETMREKVLASSARQRLRCASIQ
KFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKECCHGD
LLECADDRADLAKYICDNQDTISSKLKECCDKPLLEKSHCIAEVEK
DAIPENLPPLTADFAEDKDVCKNYQEAK**DAFLGSFLYEYSR**RHPEY
AVSVLLRLAKEYEATLEECCAKDDPHACYSTVFDKLKHLVDEPQNL
IKQNCDQFEKLGEYGFQNALIVRYTRKVPQVSTPTLVEVSRSLGKV
GTRCCTKPESERMPCTEDYLSLILNRLCVLHEKTPVSEKVTKCCTE
SLVNRRPCFSALTPDETYVPKAFDEKLFTFHADICTLPDTEKQIKKQ
TALVELLKHKPKATEEQLKTVMENFVAFVDKCCAADDKEACFAVEG
PKLVVSTQTALA

adding a chemical labeling reagent that is either covalently attached to some substrate by the target enzyme, or is converted to a lower molecular weight product. These types of proteomics experiments have been used to detect metabolic disorders in newborns (see citations).

There are many challenges to proteomic analysis. For example, the proteins present at low levels in a cell lysate can be difficult to detect. In some eukaryotic cells, the concentration differences between the most- and least-abundant proteins can be 10^6-fold, and many proteins that are interesting targets (e.g., for drug development) are low-abundance proteins. For these reasons a fractionation step prior to mass analysis may be included to either remove highly-expressed proteins and/or increase the concentrations of low-level proteins.

References

Dunn, M. J. (2000) Studying heart disease using the proteomic approach. *Drug Discov. Today* 5:76–84.

Gavin, A.-C., et al. (2001) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147.

Goh, W. W. B., Y. H. Lee, R. M. Zubaidah, J. Jin, D. Dong, Q. Lin, M. C. M. Chung, and L. Wong (2011) Network-based pipeline for analyzing MS data: An application toward liver cancer. *J. Proteome Res.* 10:2261–2272.

Graves, P. R., and T. A. Haystead (2002) Molecular biologist's guide to proteomics. *Microbiol. Mol. Biol. Rev.* 66: 39–63.

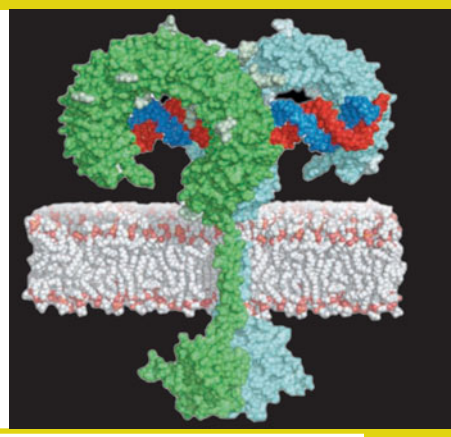
Nagaraj, S. H., R. B. Gasser, and S. Ranganathan (2006) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinform.* 8:6–21.

Ning, Z., H. Zhou, F. Wang, M. Abu-Farha, and D. Figeys (2011) Analytical aspects of proteomics: 2009–2010. *Anal. Chem.* 83:4407–4426.

Rain, J. C., L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain (2001) The protein-protein interaction map of *Helicobacter pylori*. *Nature* 409:211–215.

Spacil, Z., S. Elliott, L. Reeber, M. H. Gelb, C. R. Scott, and F. Turecek (2011) Comparative triplex tandem mass spectrometry assays of lysosomal enzyme activities in dried blood spots using fast liquid chromatography: Application to newborn screening of Pompe, Fabry, and Hurler diseases. *Anal. Chem.* 83:4822–4828.

Sutton, C. W., N. Rustogi, C. Gurkan, A. Scally, M. A. Loizidou, A. Hadjisavvas, and K. Kyriacou (2010) Quantitative proteomic profiling of matched normal and tumor breast tissues. *J. Proteome Res.* 9:3891–3902.



CHAPTER 6

The Three-Dimensional Structure of Proteins

In Chapter 5 we introduced the concept of protein primary structure. We emphasized that this first level of organization, the amino acid sequence, is dictated by the DNA sequence of the gene for the particular protein. However, nearly all proteins exhibit higher levels of structural organization as well. It is the three-dimensional structure of each protein that specifies its function in a particular biological process.

Figure 5.1 (page 137) shows a well-defined spatial location for every heavy atom within the protein sperm whale myoglobin. Figure 6.1 depicts another representation of the three-dimensional conformation of the myoglobin molecule, and illustrates that there exist two distinguishable levels of three-dimensional folding of the polypeptide chain. First, the chain appears to be locally coiled into regions of helical structure. Such local *regular* folding is called the **secondary structure** of the molecule. The helically coiled regions themselves are, in turn, folded into a specific compact structure for the entire polypeptide chain. We call this further level of folding the **tertiary structure** of the molecule. Later in this chapter we shall find that some proteins consist of several folded polypeptide chains, arranged in a regular manner. This arrangement we designate as the **quaternary** level of organization.

Protein molecules have four levels of structural organization: primary (sequence), secondary (local folding), tertiary (overall folding), and quaternary (subunit association).

This chapter is devoted to an examination of the levels of protein structure, how folded proteins are stabilized, the mechanisms by which protein folding is thought to occur, and emerging computational methods for predicting tertiary structure from primary sequence.

Secondary Structure: Regular Ways to Fold the Polypeptide Chain

Theoretical Descriptions of Regular Polypeptide Structures

Our understanding of the protein secondary structure had its origins in the remarkable work of Linus Pauling, one of the greatest chemists of the twentieth century. As early as the 1930s, he had begun X-ray diffraction studies of amino

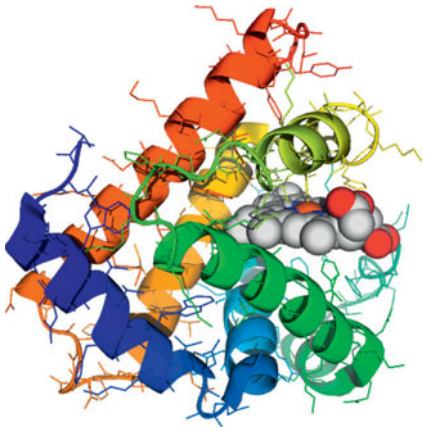


FIGURE 6.1

Three-dimensional folding of the protein myoglobin. This "cartoon" rendering was generated from the X-ray crystal structure determined by H. C. Watson and J. C. Kendrew (PDB ID: 1MBN), and it shows the polypeptide main chain as helices connected by thick lines. Side chains are shown as thin lines. Individual helical regions are color-coded, with the peptide N-terminus shown in blue and the C-terminus shown in red. This protein binds a heme group (shown in space-filling display). The orientation of the protein in this figure is the same as that shown in Figure 5.1.

acids and small peptides, with the aim of eventually analyzing protein structure. In the early 1950s, Pauling and his collaborators used these data together with remarkable scientific intuition to begin a systematic analysis of the possible regular conformations of the polypeptide chain. They postulated several principles that any such structure must obey:

1. The bond lengths and bond angles should be distorted as little as possible from those found through X-ray diffraction studies of amino acids and peptides, as shown in Figure 5.14b (page 148).
2. No two atoms should approach one another more closely than is allowed by their van der Waals radii.
3. The amide group must remain planar and in the *trans* configuration, as shown in Figure 5.14b. (This feature had been recognized in the earlier X-ray diffraction studies of small peptides.) Consequently, rotation is possible only about the two bonds adjacent to the α-carbon in each amino acid residue, as shown in Figure 6.2.
4. Some kind of noncovalent bonding is necessary to stabilize a regular folding. The most obvious possibility is hydrogen bonding between amide protons and carbonyl oxygens:

Such a concept was familiar to Pauling, who had had much to do with the development of the idea of H-bonds. In summary, the preferred conformations must be those that allow a maximum amount of hydrogen bonding, yet satisfy criteria 1–3.

Of the several possible secondary structures for polypeptides, the most frequently observed are the α helix and the β sheet.

α Helices and β Sheets

Working mainly with molecular models, Pauling and his associates were able to arrive at a small number of regular conformations that satisfied all of these criteria. Some were helical structures formed by a single polypeptide chain, and some were sheet-like structures formed by adjacent chains. The two structures they proposed as most likely—the right-handed **α helix**, and the **β sheet**—are shown in Figure 6.3a and b. These two structures are, in fact, the most commonly observed secondary structures in proteins. Figure 6.3c shows the so-called **3_{10} helix**, which is observed in some proteins but is not as common as the α helix. All of the protein secondary structures shown in Figure 6.3 satisfy the criteria listed earlier. In particular, in each structure the peptide group is planar, and every amide proton and every carbonyl oxygen (except a few near the ends of helices) is involved in hydrogen bonding. The arrangement of the main chain H-bonds in the α helix along the helical axis orients the amide N—H and C=O such that the dipole moments for each of these polar bonds align and gives rise to a **helical dipole moment** (also called a "macrodipole"). In effect, the

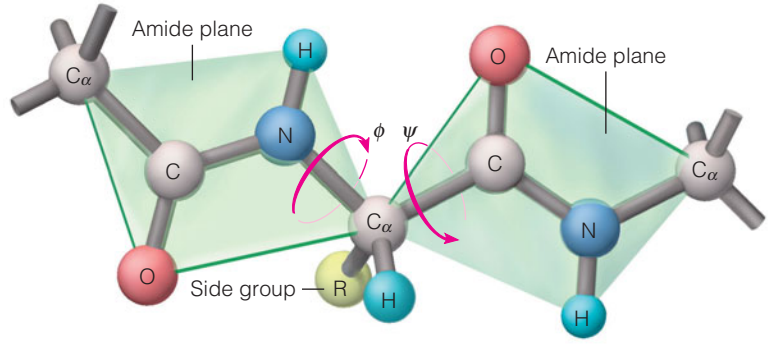
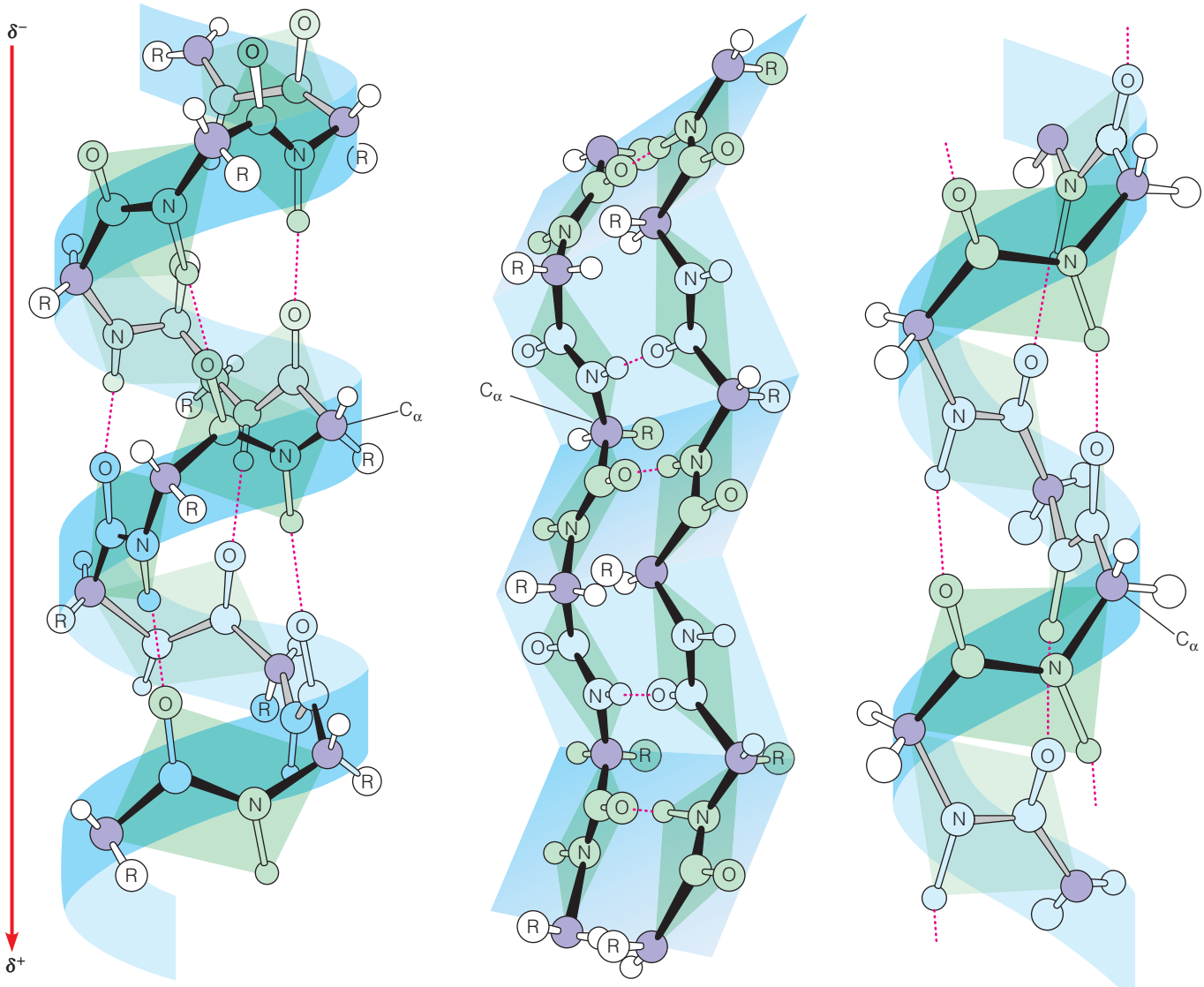


FIGURE 6.2

Rotation around the bonds in a polypeptide backbone. Two adjacent amide planes are shown in light green. Rotation is allowed only about the $N_{amide}—C_\alpha$ and $C_\alpha—C_{carbonyl}$ bonds. The angles of rotation about these bonds are defined as ϕ (phi) and ψ (psi), respectively, with directions defined as positive rotation as shown by the arrows; positive rotation is clockwise as seen from the α-carbon. The extended conformation of the chain shown here corresponds to $\varphi = +180°$, $\psi = +180°$.



(a) α helix **(b)** β sheet **(c)** 3_{10} helix

FIGURE 6.3

The right-handed α helix, β sheet, and 3_{10} helix. The right-handed α helix and β sheet are the two most frequently observed regular secondary structures of polypeptides. **(a)** In the α helix the hydrogen bonds (red dotted lines) are within a single polypeptide chain and are almost parallel to the helix axis. The alignment of amide bonds in the helix gives rise to a helical macrodipole moment shown by the red arrow (see Figure 2.4, page 29). The N-terminal end of the helix has partial (+) charge character and the C-terminal end has partial (−) charge character. **(b)** In the β sheet, the hydrogen bonds are between adjacent chains, of which only two are shown here. In this structure, the hydrogen bonds are nearly perpendicular to the chains. **(c)** The 3_{10} helix is found in proteins, but is less common than the α helix.

N-terminus of the helix has partial (+) charge character and the C-terminus has partial (−) charge character as shown by the red arrow in Figure 6.3.

Amphiphilic Helices and Sheets

In an α helix the side chains are pointing away from the center of the helix (Figure 6.4). In a β sheet, a network of main chain H-bonds connects the β strands. If we consider the H-bonded main chains to be a "sheet," the side chains are located on opposite faces of this sheet, as shown in the right-hand panels of Figure 6.4. Side chains of similar polarity are frequently clustered together to form extended hydrophilic or hydrophobic surfaces, or "faces," on one side of a helix or sheet. Secondary structures that display a predominantly hydrophobic face opposite a predominantly hydrophilic face are said to be "**amphiphilic**" (or "**amphipathic**"). Many α helices and β sheets have this characteristic because it allows two or more secondary structures to associate via contacts between the hydrophobic surfaces, while projecting the hydrophilic surfaces toward the aqueous solvent. An amphiphilic α helix will have side chains of similar polarity every 3–4 residues, whereas a β strand in an amphiphilic β sheet will have alternating polar and nonpolar side chains. These distinct patterns of side chain polarity are the basis for many secondary structure prediction algorithms (discussed below).

An amphiphilic α helix will have side chains of similar polarity every 3–4 residues, whereas a β strand in an amphiphilic β sheet will have alternating polar and nonpolar side chains.

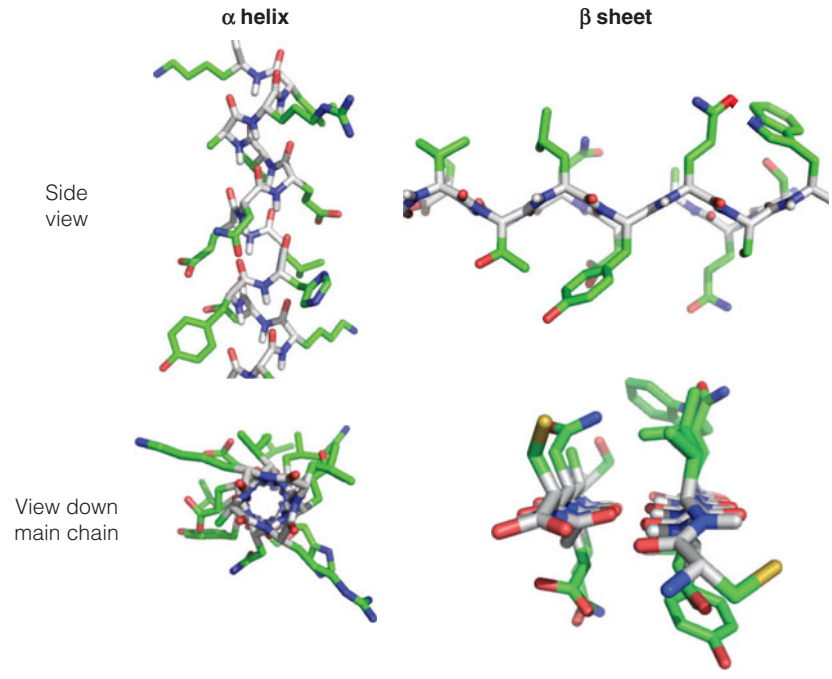


FIGURE 6.4

The positions of side chains in the α helix and β sheet. An idealized right-handed α helix and a two-stranded β sheet are shown with the main chain atoms colored gray/red/blue and side chain atoms colored green, red, blue, yellow. In the α helix the side chains radiate away from the center of the helix (lower left panel). In the β sheet the side chains are located on opposite faces of the sheet defined by the H-bonded main chain amides. This is shown in the upper right panel, looking at the main chain strands from the side (one strand is hidden behind the other), and the lower right panel, looking down the main chains of both strands.

Describing the Structures: Helices and Sheets

In Tools of Biochemistry 4A, we listed the distances that define a molecular helix: the crystallographic repeat (c), the pitch (p), and the rise (h). We also pointed out that helices may be either right-handed or left-handed and may contain either an integral number of residues per turn or a nonintegral number. We call the number of residues per turn n. Some idealized helices with integral values of n are illustrated schematically in Figure 6.5. Note that as the number of residues per turn decreases, the structure changes progressively from a broad helix to a flat ribbon ($n = 2$). Not all of these proposed helical structures are found in polypeptides. For example, the single-chain $n = 2$ structure shown in Figure 6.5 has not yet been observed in proteins.

One of Pauling's major insights was to recognize that polypeptide helices are not required to have an integral number of residues per turn. For example, the α helix repeats after exactly 18 residues, which amounts to 5 turns. It has, therefore, 3.6 residues per turn. Because the pitch of a helix is given by $p = nh$, we have for the α helix, with a rise of 0.15 nm/residue, $p = (3.6\ \text{res/turn}) \times (0.15\ \text{nm/res}) = 0.54$ nm/turn. Parameters for the secondary structures shown in Figures 6.3 and 6.4 are listed in Table 6.1.

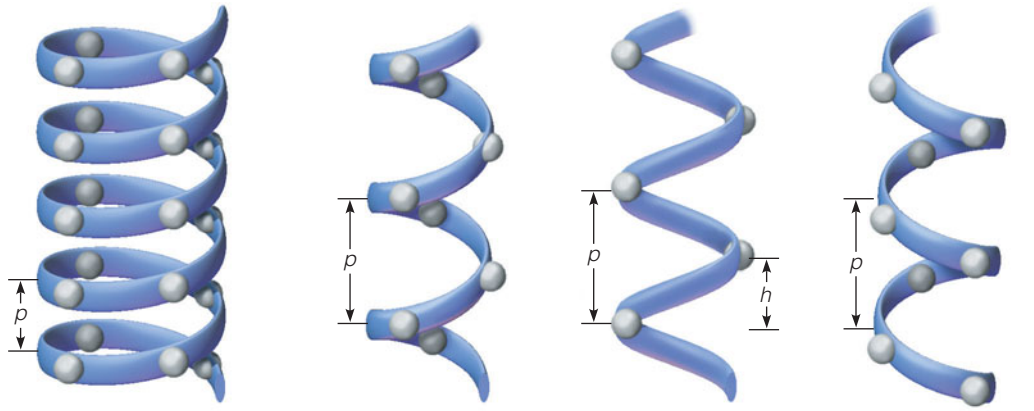


FIGURE 6.5

Idealized helices. These hypothetical structures show the effect of varying the number (n) of polypeptide residues per turn of a helix. The white balls represent α-carbon atoms. In each case the pitch (p) is indicated, and for $n = 2$ the rise (h) is also shown. The $n = 4$ and $n = 3$ helices are right-handed, the $n = -3$ helix is left-handed, and $n = 2$ (a flat ribbon) has no handedness. The right-handed α helix (not shown here), with $n = 3.6$, is intermediate between the $n = 3$ and $n = 4$ structures.

When you examine the model for the α helix (Figure 6.3a), you will note that a given carbonyl oxygen, on residue ***i***, is hydrogen-bonded to the amido proton that is four residues removed in the direction of the C-terminus (i.e., on residue $i + 4$). Thus, if we include the hydrogen bond, a loop of 13 atoms is formed. Figure 6.6 shows this schematically for the α and 3_{10} helices. Each helix type has a different number of atoms in such a hydrogen-bonded loop. We shall call this number N. Rather than using the parameters n, h, and p, an alternative way to describe a polypeptide helix is to combine n and N in the shorthand n_N. The 3_{10} helix fits this description; it has exactly 3.0 residues per turn and a 10-member hydrogen-bonded loop. The α helix could also be called a 3.6_{13} helix.

Because hydrogen bonds tend to be linear, the atoms N—H · · · O in polypeptide helices should lie on a straight line. Figure 6.3 shows that this requirement is approximately satisfied for the 3_{10} and α helices. On the other hand, it is very difficult to make helices with only two residues per turn *and* linear hydrogen bonds between residues in the same chain; thus, the only $n = 2$ structure that is found in proteins is *not* the flat ribbon shown in Figure 6.5 but the β sheet structure shown in Figure 6.3b.

A β sheet is composed of two or more ***β* strands**. Each residue in the strand is rotated by 180° with respect to the preceding one, which makes each β strand an $n = 2$ "helix." If the chains are also folded in the pleated fashion shown in Figure 6.3b, hydrogen bonds can occur *between* adjacent β strands. Forming interchain bonds allows correct bond angles with minimal strain when $n = 2$. There are two ways in which β strands can be oriented in a β sheet. The β sheet shown in Figure 6.3b shows two β strands arranged such that the N-terminus to C-terminus orientations of the two strands are in opposite directions. Such an arrangement of strands is called "antiparallel," whereas the arrangement with both strands oriented in the same direction is called "parallel" (see Figure 6.7). The hydrogen bonds between antiparallel strands are linear, whereas those between parallel strands are not.

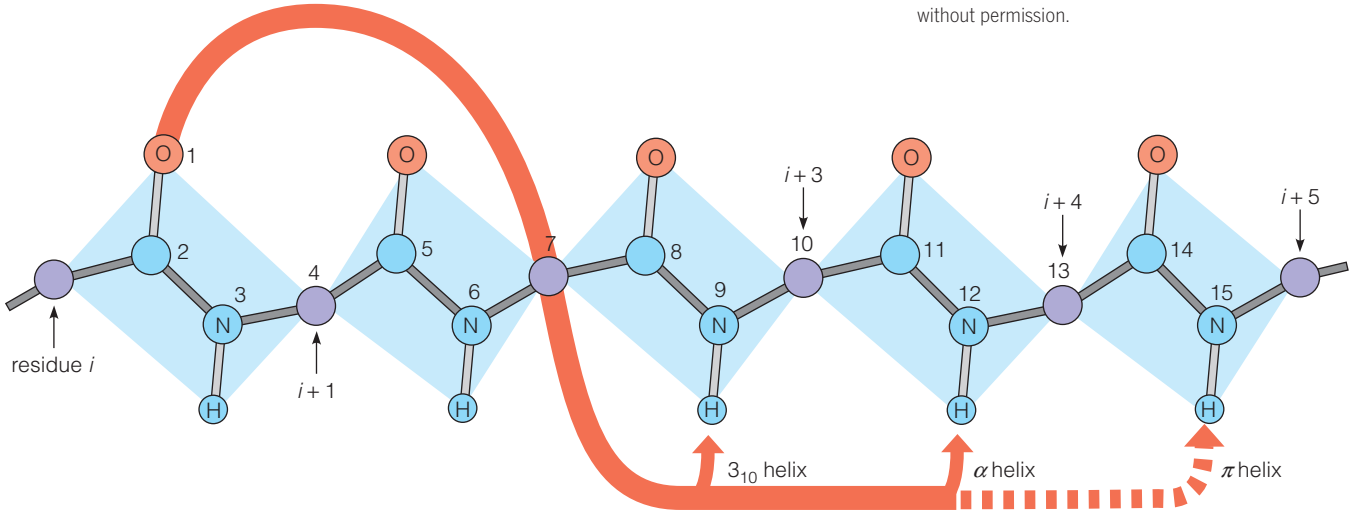
In addition to the helices and sheets described above, there is one more regularly repeating conformation that is commonly observed in protein structures—the so-called **polyproline II helix**. This particular conformation was not predicted on theoretical grounds because it does not satisfy Pauling's requirement for H-bonding. Nevertheless, it is a common motif in protein structures. Unlike the α and 3_{10} helices, this structure does not have stabilizing H-bonds between mainchain groups, and it is left-handed. Roughly a third of the amino acid residues found in this conformation are prolines, leading to the designation "polyproline II helix"; however, glycine is often found in this conformation, as are, albeit to a much lesser extent, several other amino acids. Because this conformation is not restricted to proline residues, we will refer to this secondary structure motif by the more general term "polypeptide II helix" (Figure 6.8).

TABLE 6.1 Parameters of some polypeptide secondary structures

Structure Type	Residues/ Turn	Rise (h) per residue	Pitch (p)
β Strand (antiparallel)	2.0	0.34 nm	0.68 nm
β Strand (parallel)	2.0	0.32 nm	0.64 nm
α helix	3.6	0.15 nm	0.54 nm
3_{10} helix	3.0	0.20 nm	0.60 nm
Polypeptide II helix ("polyproline II helix")	3.0	0.47 nm	0.94 nm

FIGURE 6.6

Hydrogen bonding patterns for the α and 3_{10} helices. The structures are represented in a diagrammatic way to simplify counting the atoms in each H-bonded loop. For example, there are 13 atoms in the H-bonded loop corresponding to the α (3.6_{13}) helix. The significance of the π helix is discussed later in this chapter.



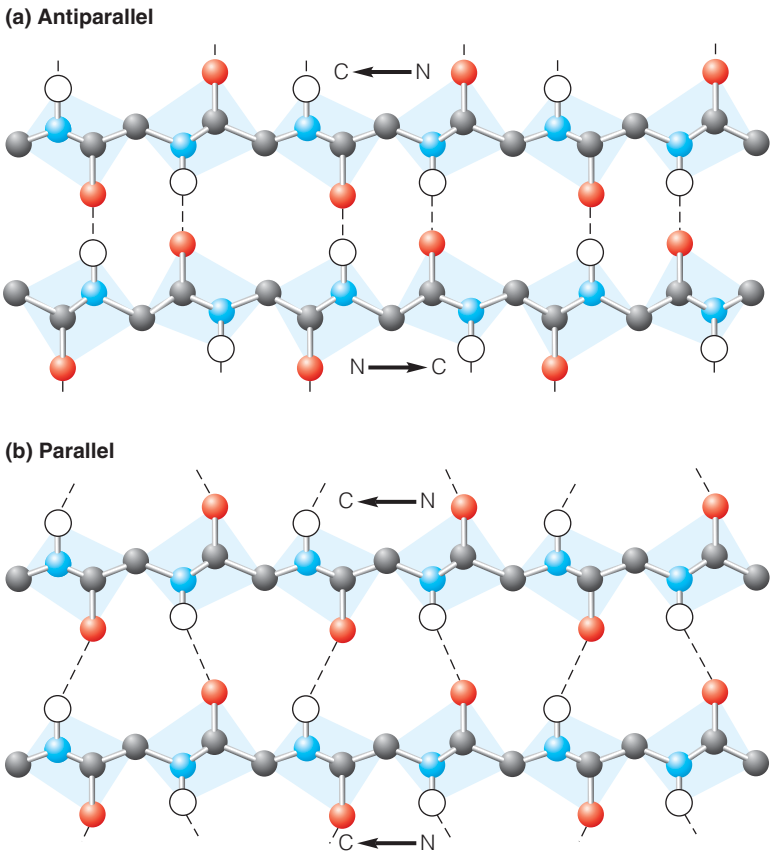


FIGURE 6.7

β Sheets. (a) An antiparallel arrangement of β strands. **(b)** A parallel arrangement of β strands. Only main chain atoms are shown (side chains omitted for clarity); H-bonds between strands are represented by dashed lines.

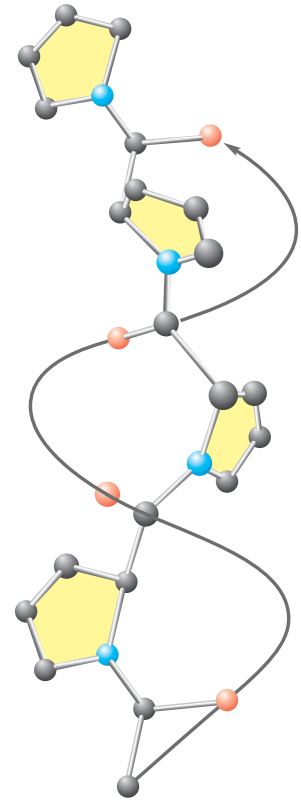


FIGURE 6.8

The polypeptide II helix. A polyproline sequence is shown; however, polyglycine also adopts this conformation. The left-handed helical twist is indicated by the curved gray arrow.

To close this section we describe a conformation called the **π helix**. The π helix is also known as an "α bulge" or "α aneurism" or "π bulge." It is widespread, as it is found in ~15% of protein sequences in the Protein Data Bank, although it is infrequent, as it typically occurs only once in any given sequence. Most (~85%) π helix conformations appear to be the result of a mutation event that results in the insertion of an amino acid into an α helix (Figure 6.9). This disrupts the normal H-bonding in the α helix, and two or more residues form H-bonds with the $i + 5$ residue, rather than the $i + 4$ residue (see Figure 6.6), creating a bulge in the helical structure. The short stretch of π helix shown in Figure 6.9 is typical of π helices found in most proteins—most are only one turn in length, and extended π helices of greater than two turns are not observed in proteins. For this reason, we will not consider the π helix as a regular repeating conformation. Nevertheless, the π helix is noteworthy because the inserted amino acid frequently confers some new functional properties on the resulting protein; thus, it is a potential marker for tracing the evolution of protein function.

We have described the common secondary structural motifs found in proteins and some of the reasons we should expect to see these structures based on consideration of (1) the planarity of the amide bond and (2) steric restrictions to rotation

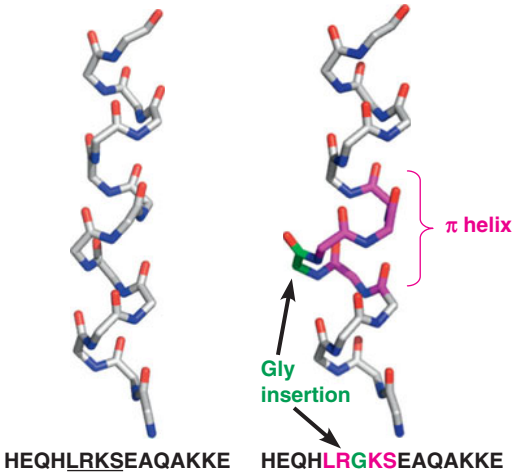


FIGURE 6.9

The π helix conformation. On the left, a main-chain rendering of the C-terminal α helix from *Staphylococcus aureus* nuclease is shown (PDB ID: 1EYD). The amino acid sequence of this helix is shown below in one-letter code. The site of Gly insertion is underlined. On the right, the analogous helix from a Gly insertion mutant is shown (PDB ID: 1STY). The inserted Gly is highlighted in green and the adjacent four residues that adopt the π helix conformation are highlighted in magenta. Note that the inserted Gly carbonyl does not form an intrahelical H-bond.

around the angles ϕ and ψ. We have also described two ways to specify the regular repeating unit of the motif– by n_N, or by listing the parameters in Table 6.1. Yet another way to describe the regular repeat of a secondary structural motif is to specify the values of the angles ϕ and ψ. As illustrated in Figure 6.10a, some combinations of ϕ and ψ (e.g., $\phi = 0°$ and $\psi = 0°$) are not allowed due to steric crowding. These steric constraints on peptide conformation can be appreciated by considering space-filling models for helices and sheets. For example, as shown in Figure 6.10b, the main chain atoms in the α helix are closely packed, with R groups projecting away from the helical axis. The combinations of ϕ and ψ angles that are most favorable (or "allowed")—because they relieve steric crowding—are shown in a systematic description of polypeptide backbone conformation called a **Ramachandran plot**.

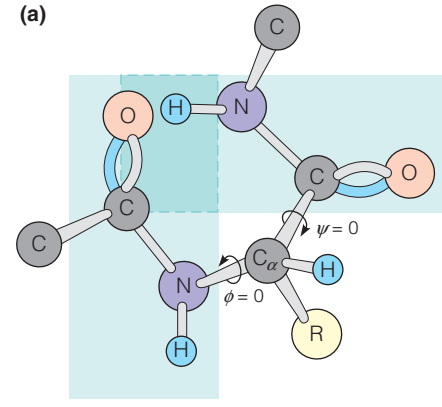
Ramachandran Plots

As is shown in Figure 6.2, each residue in a polypeptide chain has two backbone bonds about which rotation is permitted. The angles of rotation about these bonds, defined as ϕ and ψ, describe the backbone conformation of any particular residue in any protein. To make the definition meaningful, we must specify what we mean by a positive direction of rotation and the zero-angle conformation of each. The conventions chosen for directions of positive rotation about ϕ and ψ are given by the arrows in Figure 6.2—that is, clockwise when looking in either direction from the α-carbon. The conformation shown in that figure corresponds to $\phi = +180°$ and $\psi = +180°$, the fully extended form of the polypeptide chain.

With these conventions, the backbone conformation of any particular residue in a protein can be described by a point on a map (Figure 6.11) with coordinates ϕ and ψ. Such maps are called Ramachandran plots, after the biochemist G. N. Ramachandran, who first made extensive use of them. For any regular

FIGURE 6.10

Steric interactions determine peptide conformation. **(a)** A sterically nonallowed conformation. The conformation $\phi = 0°$, $\psi = 0°$ is not allowed in any polypeptide chain because of the steric crowding between the carbonyl oxygen and amido proton. **(b)** The atoms in a helix are closely packed. Here, a segment of an α helix in sperm whale myoglobin (the longer green helix in Figure 6.1; PDB ID: 1MBN) is shown as a space-filling model.



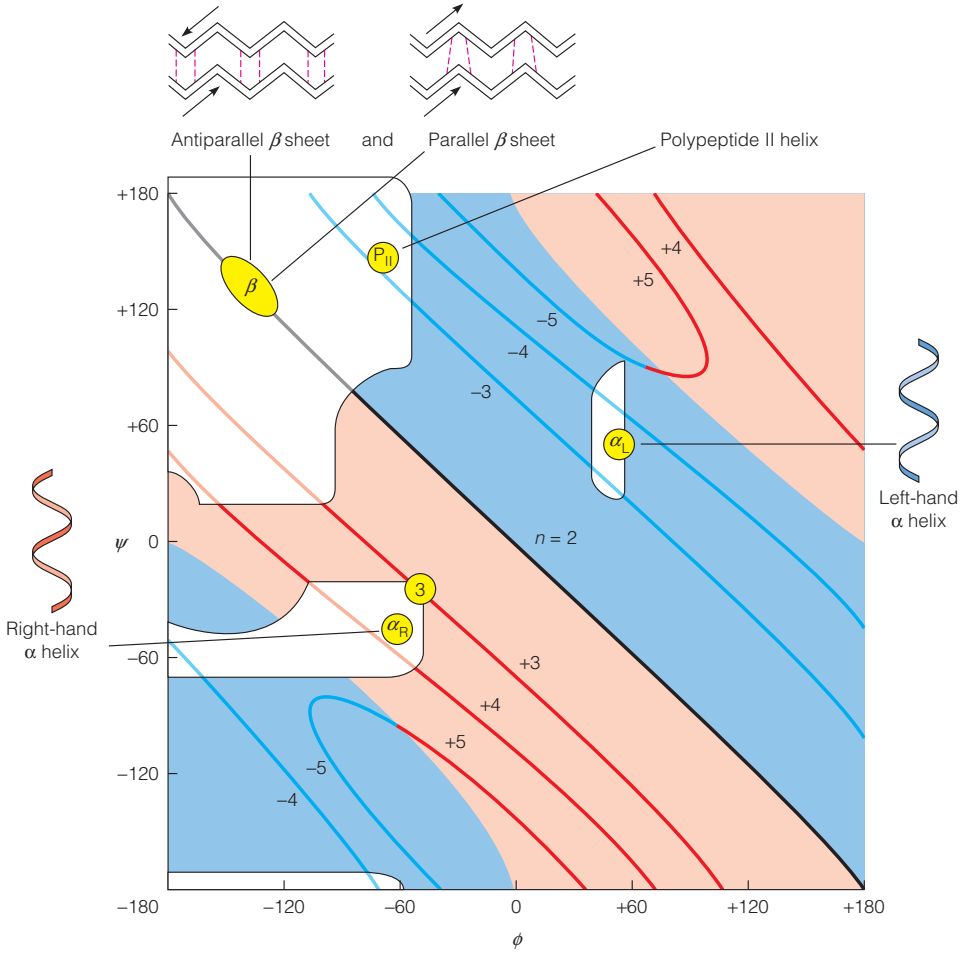


FIGURE 6.11

A Ramachandran plot for poly-L-alanine. A map of this type can be used to describe the backbone conformation of any polypeptide residue as well as the secondary structures of proteins. The coordinates are the values of the bond angles ϕ and ψ, defined as in Figure 6.2. The white areas correspond to sterically allowed conformations for poly-L-alanine (i.e., a peptide consisting of only L-Ala residues). The colored lines running across the graph correspond to various values of n (residues per turn); the line bisecting the graph corresponds to $n = 2$. The helix is right-handed when n is positive (pink and red regions), left-handed when n is negative (blue regions). The circles with the following symbols correspond to the secondary structures discussed in the text: α_R, right-handed α helix; 3, 3_{10} helix; β, β sheet; P_{II}, polypeptide II helix; α_L, left-handed α helix.

TABLE 6.2 Ranges of allowed ϕ and ψ angles for some polypeptide secondary structures

Structure Type	Φ	Ψ
β strand	$-150°$ to $-100°$	$+120°$ to $+160°$
α helix	$-70°$ to $-60°$	$-50°$ to $-40°$
3_{10} helix	$-70°$ to $-60°$	$-30°$ to $-10°$
Polypeptide II helix ("polyproline II helix")	$-80°$ to $-60°$	$+130°$ to $+160°$

Data from *Protein Science* 18:1321–1325 (2009), S. A. Hollingsworth, D. S. Berkholz, and P. A. Karplus, On the occurrence of linear groups in proteins.

Many secondary structural motifs, defined by regular repeats of ϕ and ψ angles, can be imagined; but, only a few are sterically allowed. The Ramachandran plot illustrates which combinations of ϕ and ψ angles are allowed.

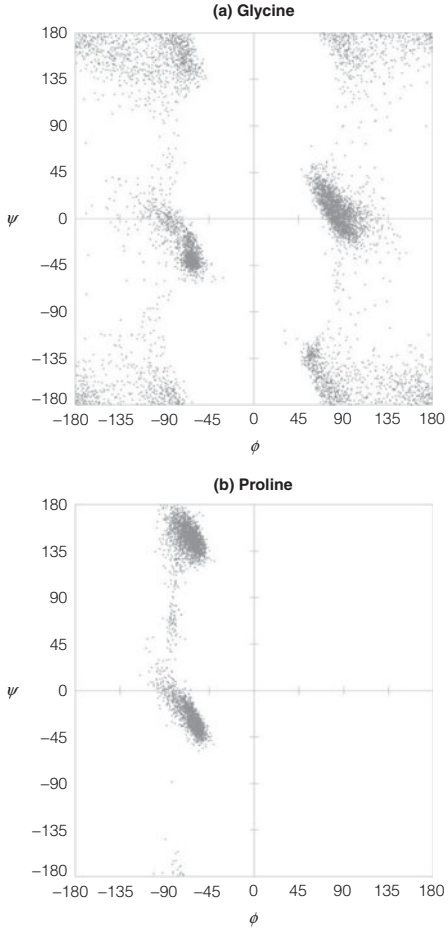


FIGURE 6.12

Ramachandran plots for glycine and proline. The data shown are for glycine and proline residues found in high-resolution X-ray crystal structures of proteins. The favorable combinations of the bond angles ϕ and ψ are shown by the darker regions. Glycine has the greatest number of allowed ϕ, ψ angle combinations, whereas proline has the fewest.

Plots courtesy of S. A. Hollingsworth and P. A. Karplus, Oregon State University.

repeating secondary structure (e.g., α helix, β sheet, etc.), all the residues that are part of the structure are in nearly equivalent conformations and have nearly equivalent ϕ and ψ angles; thus, the points on a Ramachandran plot that correspond to those residues will cluster within a narrow range of sterically allowed ϕ and ψ angles for a given secondary structure. Table 6.2 lists the ranges of ϕ and ψ angles that correspond to the various helices and β sheets described above.

One of the most useful features of Ramachandran plots is that they allow us to describe very simply which structures are sterically possible and which are not. For many of the conceivable combinations of ϕ and ψ values, some atoms in the chain would approach closer than their van der Waals radii would allow (an example is shown in Figure 6.10). Such conformations are unfavorable because they are sterically crowded. Ramachandran and other researchers have examined the entire map surface, using models and computers to determine which conformations are actually allowed. The allowed ϕ, ψ combinations for poly-L-alanine correspond to the white regions in Figure 6.11. Clearly, only a small fraction of the conceivable conformations is actually favorable. All of the regular secondary structures we have discussed fall within or very close to these regions.

Although Figure 6.11 shows the left-handed α helix lying on the edge of an allowed region, it is, in fact, not nearly as favored as the right-handed form. This difference is a consequence of the fact that all amino acids in proteins are of the L-form. With L-amino acids, steric interference between the side chains and the backbone of the helix is less with a right-handed helix than with a left-handed helix. This principle can be understood from a careful inspection of Figure 6.3a. Note that each R group is approximately *trans* to the adjacent carbonyl oxygen. If the amino acid were D instead of L, the orientation would be *cis,* with more likelihood of steric crowding. Recall from Chapter 5 that chemists have synthesized proteins with all D-amino acids. These proteins have, as expected, left-handed α helices. The importance of such side-chain effects depends on the bulkiness of the side chain. The map shown in Figure 6.11 was drawn with the assumption that all residues are L-alanine (that is, all have CH_3 side chains). If bulkier side chains were considered, the "allowed" region would shrink. Conversely, glycine, with its —H side chain, allows more conformations, as shown in Figure 6.12a. Proline has fewer combinations of allowed ϕ and ψ angles due to restricted rotation around ϕ (Figure 6.12b).

The foregoing analysis of protein structure is based on a relatively simple consideration of probable steric interactions. How does it compare with observations of actual protein structures? In this case, the correspondence between theory and observation is remarkably good. Figure 6.13 shows a plot of ϕ, ψ pairs for 30,692 residues found in 209 different polypeptide chains for which high resolution ($\leq$0.12 nm) X-ray crystallography data exist. As is evident in this figure, the majority of the observed ϕ, ψ pairs (gray dots in Figure 6.13) cluster in the regions of the Ramachandran plot that are predicted to give the most favorable combinations of ϕ and ψ (white spaces in Figure 6.11). Those residues that are classified by H-bonding patterns and geometry to be helix or sheet are identified by color in Figure 6.13 and illustrate the range of ϕ and ψ values that correspond to each of the secondary structure types discussed above (see Table 6.2).

Most of the points on this Ramachandran plot fall close to the right-hand α helix or the β sheet positions; but, they do not correspond exactly to these points, testifying to the existence of distortions of these structures in folded proteins, and to the existence of regions of structure different from either β sheet or α helix. Although most of the points fall in "allowed" regions, a few lie in "nonallowed" regions. These are mainly glycines, for which a much wider range of ϕ and ψ angles is allowed because the side chain is so small (Figure 6.12a).

Historically, the ideal and observed values of ϕ and ψ angles for parallel and antiparallel β sheets have been listed as distinct. That distinction is no longer apparent in the high-resolution structural data shown in Figure 6.13. Because the observed ϕ and ψ angles for these two β structure types overlap significantly (due to distortion of ideal angles in actual protein structures), we have chosen to list in Table 6.2 a single range of values that describes both parallel and antiparallel β sheet conformations.

Our discussion so far provides a background for understanding the basics of protein structure. It is now time to consider some specific cases. We begin with the observation that two major classes of proteins exist. These are called *fibrous* and *globular* proteins and are distinguished by major structural differences. Let us first consider the fibrous proteins.

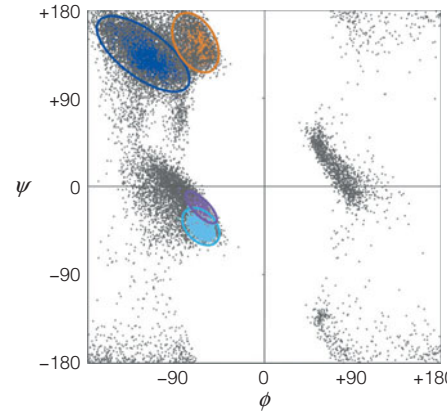


FIGURE 6.13

Observed values of ϕ and ψ from protein structural data. ϕ, ψ pairs for 30,692 residues observed in high resolution ($\leq$ 0.12 nm) crystal structures of proteins are shown (gray dots). Different colors highlight those residues defined to be right-handed α helix (cyan), right-handed 3_{10} helix (purple), β strand (blue), and left-handed polypeptide II helix (orange). The median values of ϕ and ψ for each secondary structure type are (ϕ, ψ): α helix (−63, −43), 3_{10} helix (−62, −22), β strand (−116, +129), and polypeptide II helix (−65, +145). See also Table 6.2.

Courtesy of P. A. Karplus, Oregon State University.

Fibrous Proteins: Structural Materials of Cells and Tissues

Fibrous proteins are distinguished from globular proteins by their filamentous, or elongated, form. Most of them play structural roles in animal cells and tissues—they hold things together. Fibrous proteins include the major proteins of skin and connective tissue and of animal fibers like hair and silk. The amino acid sequence of each of these proteins favors a particular kind of secondary structure, which confers on each protein a particular set of appropriate mechanical properties. Table 6.3 lists

TABLE 6.3 Amino acid compositions of some fibrous proteins

Amino Acid	α-Keratin (wool)	Fibroin (silk)	Collagen (Bovine tendon)	Elastin (Pig aorta)	All proteins[f]
Gly	8.1	44.6	32.7	32.3	7.9
Ala	5.0	29.4	12.0	23.0	8.7
Ser	10.2	12.2	3.4	1.3	5.8
Glu + Gln	12.1	1.0	7.7	2.1	6.6 (3.7)
Cys	11.2	0	0	—[e]	1.3
Pro	7.5	0.3	22.1[a]	10.7[c]	4.7
Arg	7.2	0.5	5.0	0.6	5.0
Leu	6.9	0.5	2.1	5.1	8.9
Thr	6.5	0.9	1.6	1.6	5.6
Asp + Asn	6.0	1.3	4.5	0.9	5.9 (4.2)
Val	5.1	2.2	1.8	12.1	7.2
Tyr	4.2	5.2	0.4	1.7	3.5
Ile	2.8	0.7	0.9	1.9	5.5
Phe	2.5	0.5	1.2	3.2	4.0
Lys	2.3	0.3	3.7[b]	3.6[d]	5.5
Trp	1.2	0.2	0	—[e]	1.5
His	0.7	0.2	0.3	—[e]	2.4
Met	0.5	0	0.7	—[e]	2.0

Note: The three most abundant amino acids in each protein are indicated in red. Values given are in mole percent.

[a]About 39% of this is hydroxyproline.

[b]About 14% of this is hydroxylysine.

[c]About 13% of this is hydroxyproline.

[d]Most (about 80%) is involved in cross-links.

[e]Essentially absent.

[f]Reprinted from *Journal of Chemical Information and Modeling* 50:690–700, J. M. Otaki, M. Tsutsumi, T. Gotoh, and H. Yamamoto, Secondary structure characterization based on amino acid composition and availability in proteins. © 2010 American Chemical Society.

Fibrous proteins are elongated molecules with well-defined secondary structures. They usually play structural roles in the cell.

the amino acid composition of four examples of fibrous proteins: α-keratin, fibroin, collagen, and elastin. Compared to the typical distributions of the 20 amino acids in globular proteins (see "All proteins" column in Table 6.3) each of these fibrous proteins is significantly enriched in 3–4 particular amino acids, which stabilize the extended secondary structures typical of the fibrous proteins.

The Keratins

Two important classes of proteins that have similar amino acid sequences and biological function are called α- and β-keratins. The ***α*-keratins** are the major proteins of hair and fingernails and comprise a major fraction of animal skin. The α-keratins are members of a broad group of **intermediate filament proteins**, which play important structural roles in the nuclei, cytoplasm, and surfaces of many cell types. All of the intermediate filament proteins are predominantly α-helical in structure; in fact, it was the characteristic X-ray diffraction pattern of α-keratin that Pauling and his colleagues sought to explain by their α helix model.

α-Keratin is built on a coiled-coil α-helical structure.

The structure of a typical α-keratin, that of hair, is depicted in Figure 6.14. The individual molecules contain long sequences—over 300 residues in length—that are wholly α-helical. Pairs of these right-handed helices twist about one another in a left-handed **coiled-coil** structure. This pairing of α helices appears to be a consequence of a peculiarity of the amino acid sequence of α-keratin. Every third or fourth amino acid has a nonpolar, hydrophobic side chain. Because the α helix has 3.6 residues/turn, this means there is a strip of contiguous hydrophobic surface area along one face of each helical chain (the hydrophobic surface area makes a shallow spiral around the helix because 4.0 does not exactly equal 3.6). As we noted in Chapter 2, hydrophobic surfaces tend to associate in aqueous

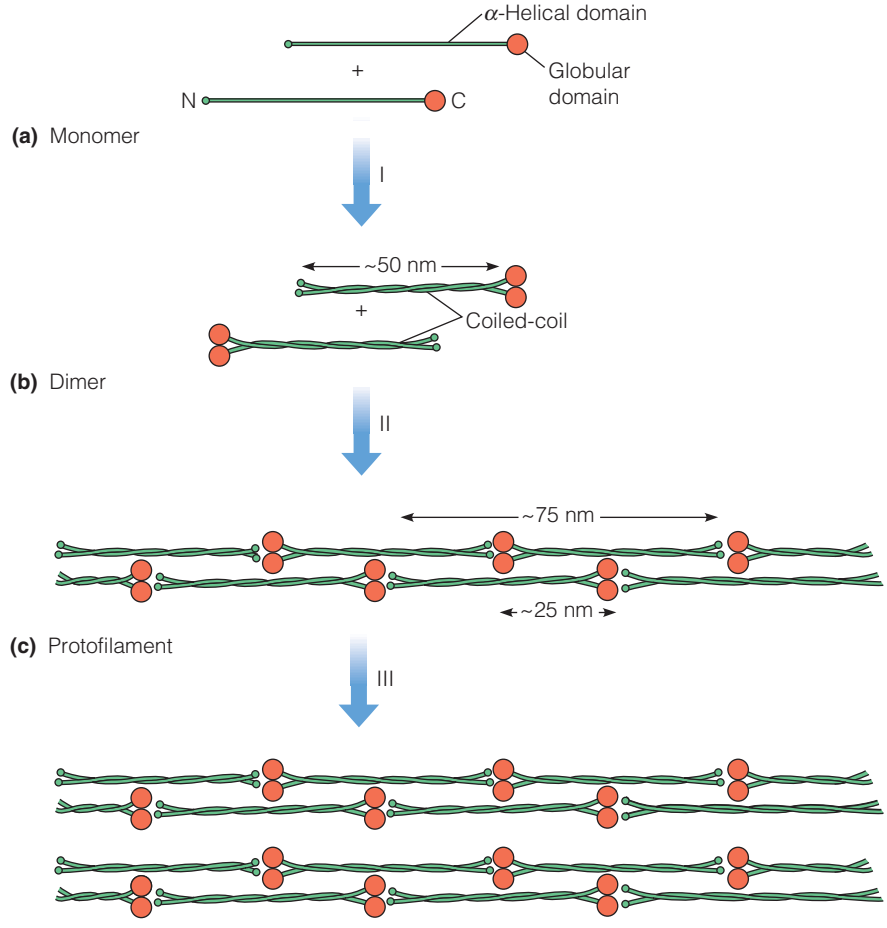


FIGURE 6.14

Proposed structure for keratin-type intermediate filaments. Two monomers **(a)** pair via a parallel coiled-coil to form the 50-nm-long dimer **(b).** These then associate to form first a 4-strand protofilament **(c)** and then an 8-strand protofibril **(d).** The regular spacing of 25 nm along the fibers is accounted for by the overlap.

medium; thus, two α-keratin helices are noncovalently bonded by hydrophobic interactions between the entwined helices.

In intermediate filaments, pairs of coiled coils themselves tend to associate into a four-chain protofilament (Figure 6.14c), and two of these in turn pack together to form a protofibril (Figure 6.14d). The details of these higher levels of association are still unclear. Such twisted cables can be very stretchy and flexible, but in different tissues α-keratin is hardened, to differing degrees, by the introduction of disulfide cross-links within the several levels of fiber structure (note that α-keratin has an unusually high content of cysteine—see Table 6.3). Fingernails have many cross-links in their α-keratin, whereas hair has relatively few. The process of introducing a "permanent wave" into human hair involves reduction of these disulfide bonds, rearrangement of the fibers, and reoxidation to "set" the tight curls thus introduced.

The β-keratins, as their name implies, contain much more β-sheet structure. Indeed, they represented the second major structural class described by Pauling and his coworkers. The β-keratins are found mostly in birds and reptiles, in structures like feathers and scales.

Fibroin

Fibroin is a β sheet protein. Almost half of its residues are glycine.

The β-sheet structure is most elegantly utilized in the fibers spun by silkworms and spiders. Silkworm fibroin (Figure 6.15) contains long regions of antiparallel β sheet, with the polypeptide chains running parallel to the fiber axis. The β-sheet regions comprise almost exclusively multiple repetitions of the sequence

$$[\,\text{Gly-Ala-Gly-Ala-Gly-Ser-Gly-Ala-Ala-Gly-(Ser-Gly-Ala-Gly-Ala-Gly)}_8\,]$$

In silkworm fibroin almost every other residue is Gly, which is usually followed by Ala or Ser residues. In other species, the residues following the Gly residues are different, leading to differences in physical properties. This alternating pattern of residues allows the sheets to fit together and pack on top of one another in the manner shown in Figure 6.15. This arrangement of sheets results in a fiber that is strong and relatively inextensible because the covalently bonded chains are stretched to nearly their maximum possible length. Yet the fibers are very flexible because bonding between the sheets involves only the weak van der Waals interactions between the side chains, which provide little resistance to bending.

Not all of the fibroin protein is in β sheets. As the amino acid composition in Table 6.3 shows, fibroin contains small amounts of other, bulky amino acids like valine and tyrosine, which would not fit into the structure shown. These are carried in compact folded regions that periodically interrupt the β-sheet segments,

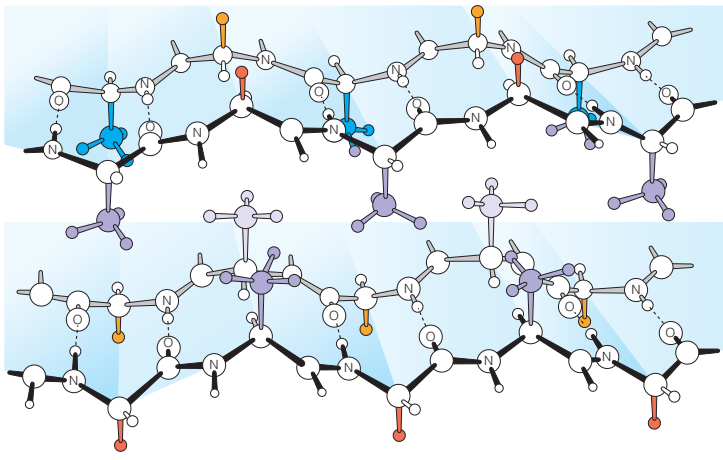


FIGURE 6.15

The structure of silk fibroin. (a) A three-dimensional view of the stacked β sheets of fibroin, with the side chains shown in color. The region shown contains only alanine and glycine residues. **(b)** Interdigitation of alanine or serine side chains and glycine side chains in fibroin. The plane of the section is perpendicular to the folded sheets.

and they probably account for the amount of stretchiness that silk fibers have. In fact, different species of silkworms produce fibroins with different extents of such non–β-sheet structure and corresponding differences in elasticity. The overall fibroin structure is a beautiful example of a protein molecule that has evolved to perform a particular function—to provide a tough, yet flexible, fiber for the silkworm's cocoon or the spider's web.

Collagen

Collagen fibers are built from triple helices of polypeptides rich in glycine and proline.

Because it performs such a wide variety of functions, **collagen** is the most abundant single protein in most vertebrates. In large animals, it may make up a third of the total protein mass. Collagen fibers form the *matrix* material in bone, on which the mineral constituents precipitate; these fibers constitute the major portion of tendons; and a network of collagen fibers is an important constituent of skin. Basically, collagen holds most animals together.

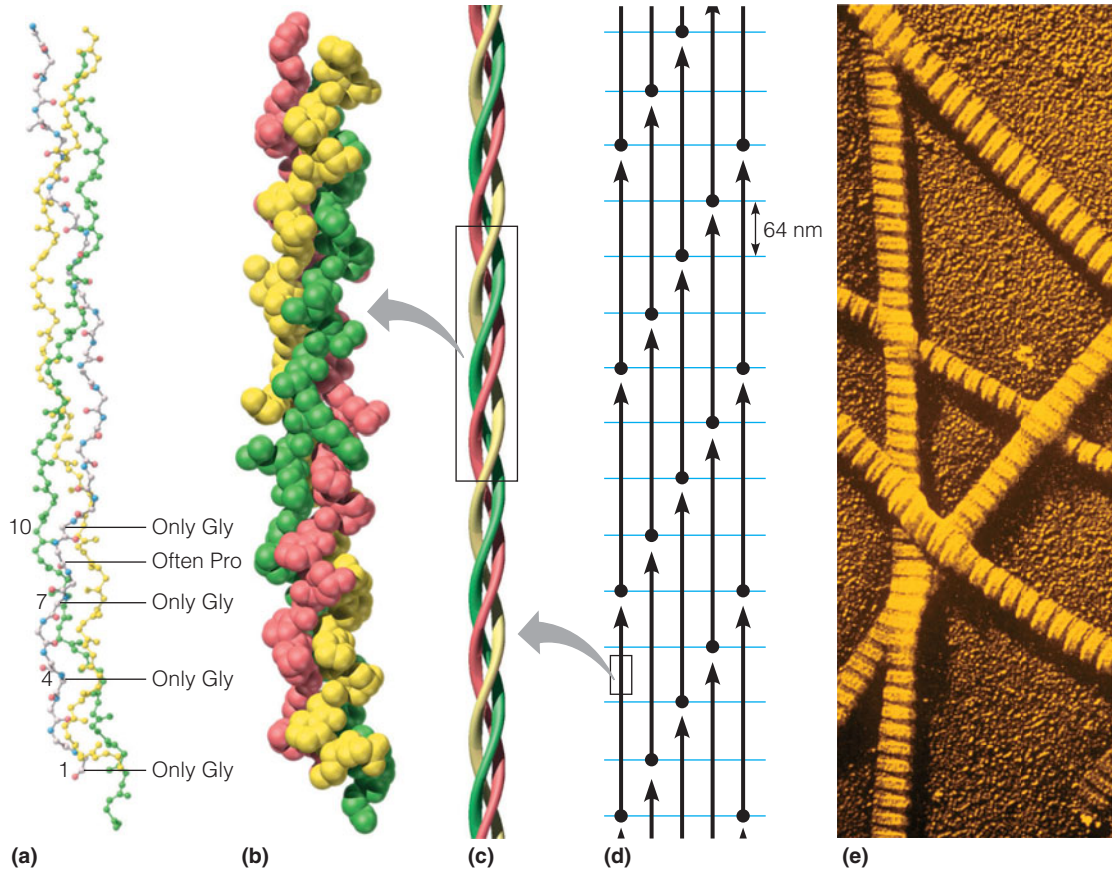
Collagen Structure

The basic unit of the collagen fiber is the **tropocollagen** molecule, a *triple helix* of three polypeptide chains, each about 1000 residues in length. This triple helical structure, shown in Figure 6.16a and b, is unique to collagen. The individual chains are *left-handed* helices, with about 3.3 residues/turn. Three of these chains wrap around one another in a right-handed sense, with hydrogen bonds extending between the chains. Examination of the model reveals that every third residue, which must lie near the center of the triple helix, can be *only* glycine (see Figure 6.16a, and Table 6.3). Any side chain other than —H would be too bulky. Formation of the individual helices of the collagen type is also favored by the presence of proline or hydroxyproline in the tropocollagen molecule. A repetitive motif in the sequence is of the form Gly–X–Y, where X is often proline and Y is

FIGURE 6.16

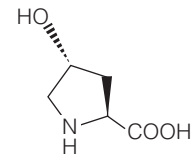
The structure of collagen fibers. The protein collagen is made up of tropocollagen molecules packed together to form fibers. The tropocollagen molecule is a triple helix. **(a)** and **(b)** Stick and space-filling views of the tropocollagen triple helix. **(c)** A lower magnification model emphasizes the interwoven triple-helical secondary structure. **(d)** Tropocollagen triple helices align side by side in a staggered fashion to form the collagen fiber. This regular arrangement leads to periodic pattern of bands separated by 64 nm (blue lines) **(e)** An electron micrograph of collagen shows the crisscrossing of fibers, with the 64 nm periodic pattern clearly visible in each.

(e) J. Gross, Biozentrum/Science Photo Library.

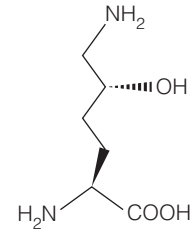


proline or hydroxyproline (see Table 6.3). However, other residues are often tolerated in these positions. Like silk fibroin, collagen is a good example of how a particular kind of repetitive sequence dictates a particular structure. In order to properly serve the multiple functions it does, collagen exists in a larger number of genetic variants in higher organisms.

Collagen is also unusual in its widespread modification of proline to hydroxyproline. Most of the hydrogen bonds between chains in the triple helix are from amide protons to carbonyl oxygens, but the —OH groups of hydroxyproline also seem to participate in stabilizing the structure. Hydroxylation of lysine residues in collagen also occurs, but is much less frequent. It plays a different role, serving to form attachment sites for polysaccharides.



(2S,4R)-4-hydroxyproline



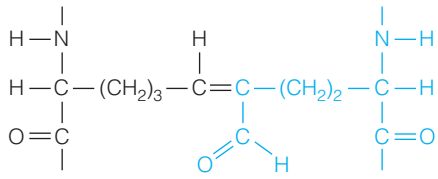
(2S,5R)-5-hydroxylysine

The enzymes that catalyze the hydroxylations of proline and lysine residues in collagen require **vitamin C**, L-ascorbic acid (see Figure 21.33, page 906). A symptom of extreme vitamin C deficiency, called **scurvy**, is the weakening of collagen fibers caused by the failure to hydroxylate these side chains, which results in reduced H-bonding between chains. Consequences are as might be expected: Lesions develop in skin and gums, and blood vessels weaken. The condition quickly improves with administration of vitamin C.

Scurvy is caused by failure to hydroxylate prolines and lysines in collagen.

The individual tropocollagen molecules pack together in a collagen fiber in a specific way (Figure 6.16c). Each molecule is about 300 nm long and overlaps its neighbor by about 64 nm, producing the characteristic banded appearance of the fibers shown in Figure 6.16e. This structure contributes remarkable strength: Collagen fibers in tendons have a strength comparable to that of hard-drawn copper wire.

Part of the toughness of collagen is due to the cross-linking of tropocollagen molecules to one another via a reaction involving lysine side chains. Some of the lysine side chains are oxidized to aldehyde derivatives, which can then react with either a lysine residue, or with one another via an aldol condensation and dehydration, to produce a cross-link:



This process continues through life, and the accumulating cross-links make the collagen steadily less elastic and more brittle. As a result, bones and tendons in older individuals are more easily snapped, and the skin loses much of its elasticity. Many of the signs we associate with aging are consequences of this simple cross-linking process.

Collagen Synthesis

As you will have judged by now, collagen is a protein that undergoes extensive modification. Indeed, it can be considered an almost complete example of the post-translational modification pathways we discussed at the end of Chapter 5. The tropocollagen triple helix that ends up cross-linked into an extracellular collagen fiber is very different from the molecule that is first synthesized on a ribosome. The steps in this transformation are shown in Figure 6.17, which begins with translation (step 1). The newly translated polypeptide is hydroxylated (step 2), and then sugars are attached to some of the newly hydroxylated lysine side chains (step 3) to yield **procollagen** (step 4). Procollagen contains almost 1500 residues, of which about 500 are in N-terminal and C-terminal regions that do not have the typical collagen fiber sequence described earlier. Three molecules of procollagen wrap their central regions into a triple helix, while the N- and C-terminal regions fold into globular protein structures. The procollagen triplexes are then exported into the extracellular space (step 5), at which point the N- and C-terminal regions

Collagen undergoes extensive post-translational modification.

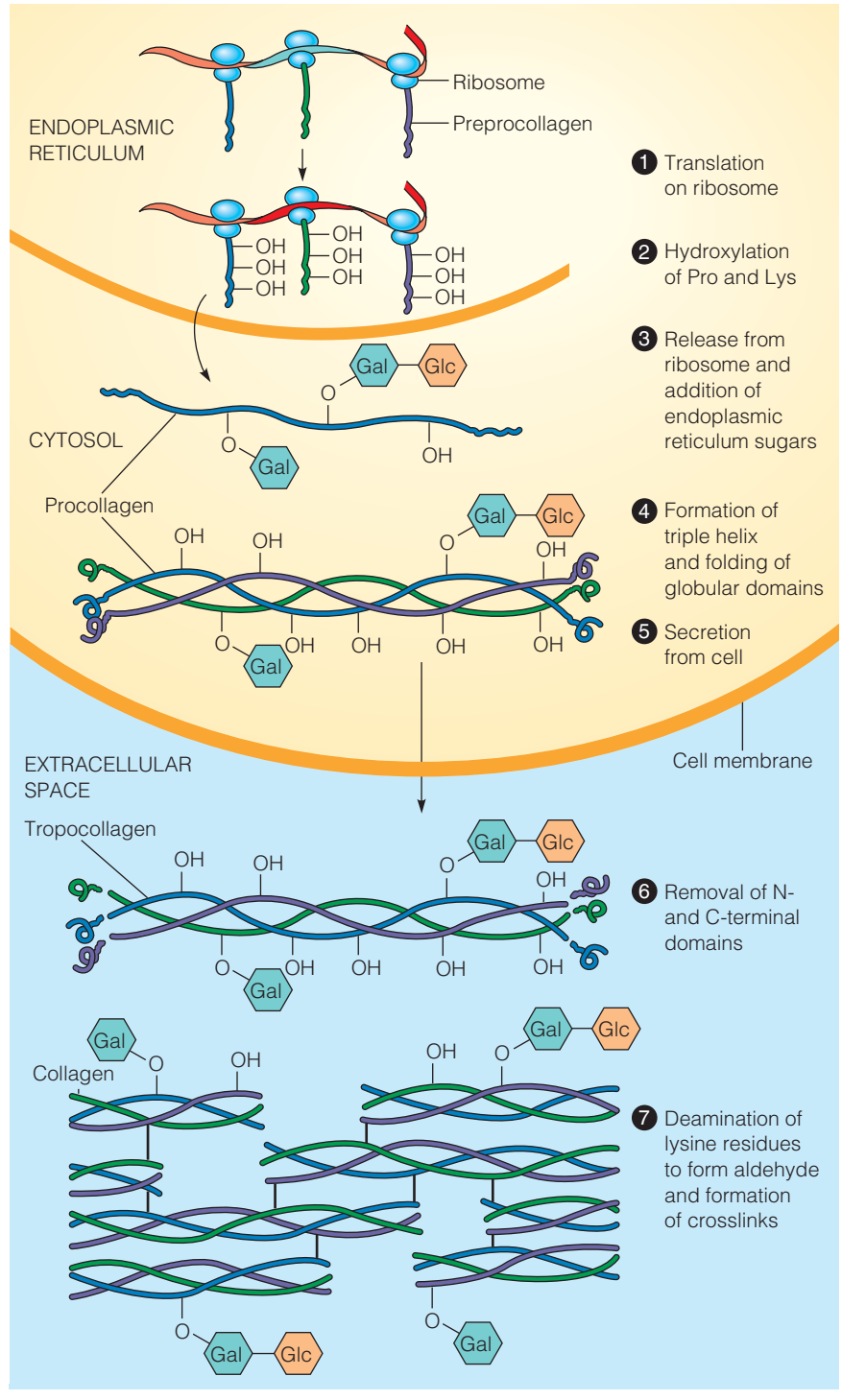


FIGURE 6.17

Biosynthesis and assembly of collagen. The process can be visualized in several steps. Steps 1–4 occur in the endoplasmic reticulum and cytosol of collagen-synthesizing cells; steps 6 and 7 occur in the extracellular region. Gal = galactose, Glc = glucose.

are cleaved off by specific proteases, leaving only the tropocollagen triple helix, about 1000 residues long (step 6). These molecules then assemble into the staggered arrays shown in Figure 6.16d. Finally, cross-linking cements the molecules together into a tough collagen fiber.

Elastin

The protein elastin forms elastic fibers found in ligaments and blood vessels.

Collagen is found in tissues where strength or toughness is required, but some tissues, such as ligaments and arterial blood vessels, need highly elastic fibers. Such tissues contain large amounts of the fibrous protein **elastin**.