BIOCHEMISTRY

FIFTH EDITION

Jeremy M. Berg
Johns Hopkins University School of Medicine

John L. Tymoczko
Carleton College

Lubert Stryer
Stanford University

Web content by
Neil D. Clarke
Johns Hopkins University School of Medicine

W. H. Freeman and Company
New York

Novo Nordisk Ex. 2076, P. 1
Mylan Institutional v. Novo Nordisk
IPR2020-00324

General Library System
University of Wisconsin - Madison
728 State Street
Madison, WI 53706-1494
U.S.A.

Steenbock Memorial Library
University of Wisconsin - Madison
550 Babcock Drive
Madison, WI 53706-1293

PUBLISHER: **Michelle Julet**

DEVELOPMENT EDITOR: **Susan Moran**

NEW MEDIA DEVELOPMENT EDITOR: **Sonia Divittorio**

NEW MEDIA AND SUPPLEMENTS EDITOR: **Mark Santee**

MEDIA DEVELOPERS: **CADRE design; molvisions.com—3D molecular visualization**

MARKETING MANAGER: **Carol Coffey**

PROJECT EDITOR: **Georgia Lee Hadler**

MANUSCRIPT EDITOR: **Patricia Zimmerman**

COVER AND TEXT DESIGN: **Victoria Tomaselli**

COVER ILLUSTRATION: **Tomo Narashima**

ILLUSTRATION COORDINATOR: **Cecilia Varas**

ILLUSTRATIONS: **Jeremy Berg with Network Graphics**

PHOTO EDITOR: **Vikii Wong**

PHOTO RESEARCHER: **Dena Betz**

PRODUCTION COORDINATOR: **Julia DeRosa**

COMPOSITION: **TechBooks**

MANUFACTURING: **RR Donnelley & Sons Company**

Library of Congress Cataloguing-in-Publication Data

Berg, Jeremy Mark
Biochemistry / Jeremy Berg, John Tymoczko, Lubert Stryer.—5th ed.
p. cm.
Fourth ed. by Lubert Stryer.
Includes bibliographical references and index.
ISBN 0-7167-4954-8 (CH1–31 only)—ISBN 0-7167-3051-0 (CH1–34)
ISBN 0-7167-4955-6 (CH32–34 only)—ISBN 0-7167-4684-0 (CH1–34, International edition)
1. Biochemistry. I. Tymoczko, John L. II. Stryer, Lubert. III. Title.

QP514.2 .S66 2001
572—dc21

2001051259

© 2002 by W. H. Freeman and Company; © 1975, 1981, 1988, 1995 by Lubert Stryer. All rights reserved

No part of this book may be reproduced by any mechanical, photographic, or electronic process, or in the form of a phonographic recording, nor may it be stored in a retrieval system, transmitted, or otherwise copied for public or private use, without written permission from the publisher.

Printed in the United States of America

First printing 2001

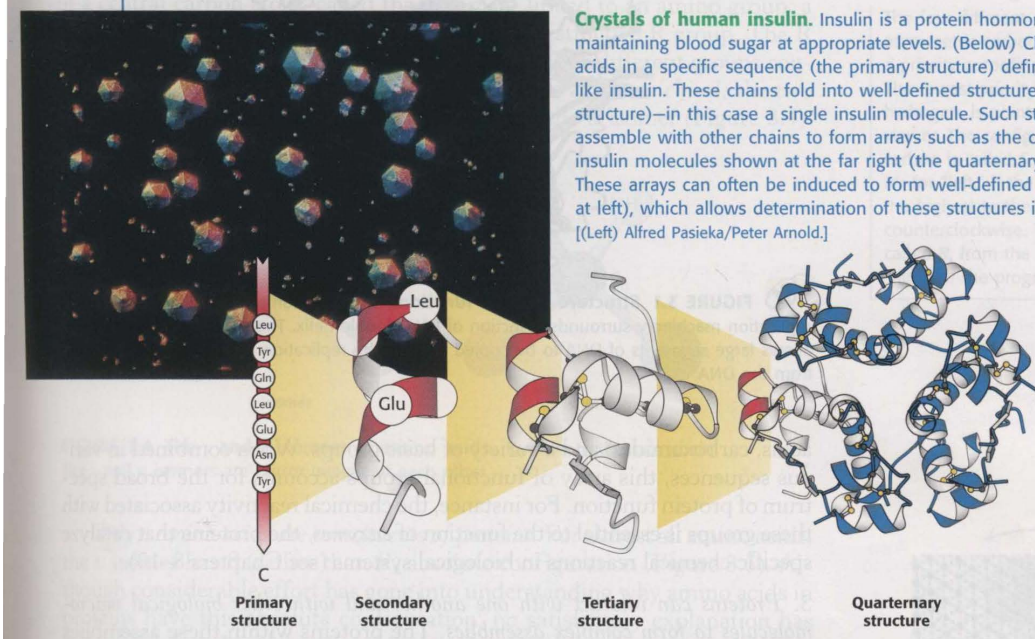
W. H. Freeman and Company
41 Madison Avenue, New York, New York 10010
Houndmills, Basingstoke RG21 6XS, England

Novo Nordisk Ex. 2076, P. 2
Mylan Institutional v. Novo Nordisk
IPR2020-00324

This material may be protected by Copyright law (Title 17 U.S. Code)

CHAPTER 3

Protein Structure and Function



Crystals of human insulin. Insulin is a protein hormone, crucial for maintaining blood sugar at appropriate levels. (Below) Chains of amino acids in a specific sequence (the primary structure) define a protein like insulin. These chains fold into well-defined structures (the tertiary structure)—in this case a single insulin molecule. Such structures assemble with other chains to form arrays such as the complex of six insulin molecules shown at the far right (the quarternary structure). These arrays can often be induced to form well-defined crystals (photo at left), which allows determination of these structures in detail. [(Left) Alfred Pasieka/Peter Arnold.]

Proteins are the most versatile macromolecules in living systems and serve crucial functions in essentially all biological processes. They function as catalysts, they transport and store other molecules such as oxygen, they provide mechanical support and immune protection, they generate movement, they transmit nerve impulses, and they control growth and differentiation. Indeed, much of this text will focus on understanding what proteins do and how they perform these functions.

Several key properties enable proteins to participate in such a wide range of functions.

1. *Proteins are linear polymers built of monomer units called amino acids.* The construction of a vast array of macromolecules from a limited number of monomer building blocks is a recurring theme in biochemistry. Does protein function depend on the linear sequence of amino acids? The function of a protein is directly dependent on its three-dimensional structure (Figure 3.1). Remarkably, proteins spontaneously fold up into three-dimensional structures that are determined by the sequence of amino acids in the protein polymer. Thus, *proteins are the embodiment of the transition from the one-dimensional world of sequences to the three-dimensional world of molecules capable of diverse activities.*

2. *Proteins contain a wide range of functional groups.* These functional groups include alcohols, thiols, thioethers, carboxylic

OUTLINE

Novo Nordisk Ex. 2076, P. 3
Mylan Institutional v. Novo Nordisk
IPR2020-00324

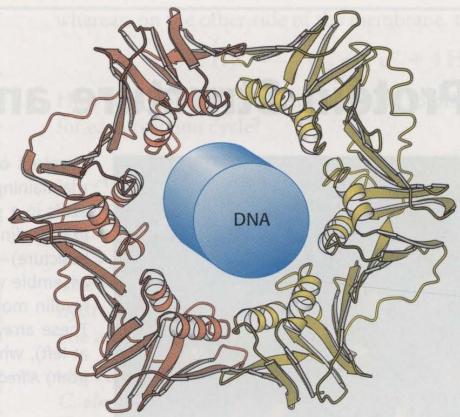


FIGURE 3.1 Structure dictates function. A protein component of the DNA replication machinery surrounds a section of DNA double helix. The structure of the protein allows large segments of DNA to be copied without the replication machinery dissociating from the DNA.

acids, carboxamides, and a variety of basic groups. When combined in various sequences, this array of functional groups accounts for the broad spectrum of protein function. For instance, the chemical reactivity associated with these groups is essential to the function of *enzymes*, the proteins that catalyze specific chemical reactions in biological systems (see Chapters 8–10).

3. *Proteins can interact with one another and with other biological macromolecules to form complex assemblies.* The proteins within these assemblies can act synergistically to generate capabilities not afforded by the individual component proteins (Figure 3.2). These assemblies include macromolecular machines that carry out the accurate replication of DNA, the transmission of signals within cells, and many other essential processes.

4. *Some proteins are quite rigid, whereas others display limited flexibility.* Rigid units can function as structural elements in the cytoskeleton (the internal scaffolding within cells) or in connective tissue. Parts of proteins with limited flexibility may act as hinges, springs, and levers that are crucial to protein function, to the assembly of proteins with one another and with other molecules into complex units, and to the transmission of information within and between cells (Figure 3.3).

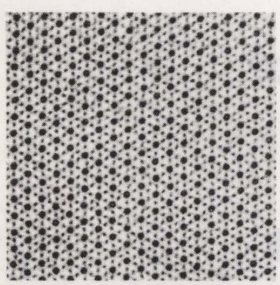


FIGURE 3.2 A complex protein assembly. An electron micrograph of insect flight tissue in cross section shows a hexagonal array of two kinds of protein filaments. [Courtesy of Dr. Michael Reedy.]

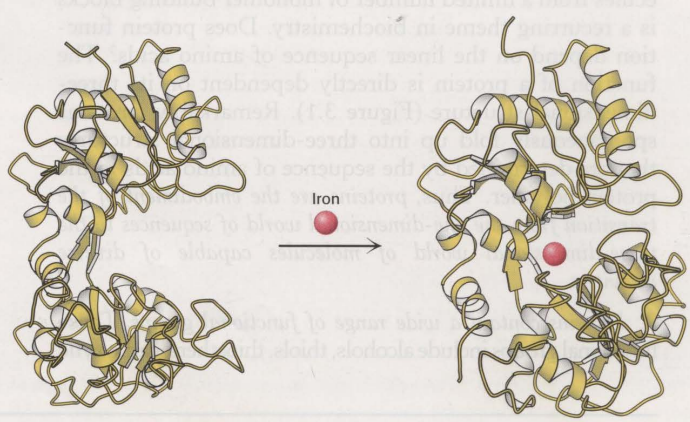


FIGURE 3.3 Flexibility and function. Upon binding iron, the protein lactoferrin undergoes conformational changes that allow other molecules to distinguish between the iron-free and the iron-bound forms.

Novo Nordisk Ex. 2076, P. 4
Mylan Institutional v. Novo Nordisk
IPR2020-00324

3.1 PROTEINS ARE BUILT FROM A REPERTOIRE OF 20 AMINO ACIDS

Amino acids are the building blocks of proteins. An *α-amino acid* consists of a central carbon atom, called the *α carbon*, linked to an amino group, a carboxylic acid group, a hydrogen atom, and a distinctive R group. The R group is often referred to as the *side chain*. With four different groups connected to the tetrahedral α-carbon atom, α-amino acids are *chiral;* the two mirror-image forms are called the L isomer and the D isomer (Figure 3.4).

Notation for distinguishing stereoisomers– The four different substituents of an asymmetric carbon atom are assigned a priority according to atomic number. The lowest-priority substituent, often hydrogen, is pointed away from the viewer. The configuration about the carbon is called *S*, from the Latin *sinister* for "left," if the progression from the highest to the lowest priority is counterclockwise. The configuration is called *R*, from the Latin *rectus* for "right," if the progression is clockwise.

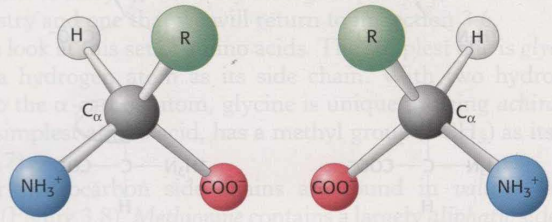


FIGURE 3.4 The L and D isomers of amino acids. R refers to the side chain. The L and D isomers are mirror images of each other.

Only L amino acids are constituents of proteins. For almost all amino acids, the L isomer has *S* (rather than *R*) absolute configuration (Figure 3.5). Although considerable effort has gone into understanding why amino acids in proteins have this absolute configuration, no satisfactory explanation has been arrived at. It seems plausible that the selection of L over D was arbitrary but, once made, was fixed early in evolutionary history.

Amino acids in solution at neutral pH exist predominantly as *dipolar ions* (also called *zwitterions*). In the dipolar form, the amino group is protonated ($-NH_3^+$) and the carboxyl group is deprotonated ($-COO^-$). The ionization state of an amino acid varies with pH (Figure 3.6). In acid solution (e.g., pH 1), the amino group is protonated ($-NH_3^+$) and the carboxyl group is not dissociated (–COOH). As the pH is raised, the carboxylic acid is the first group to give up a proton, inasmuch as its pK_a is near 2. The dipolar form persists until the pH approaches 9, when the protonated amino group

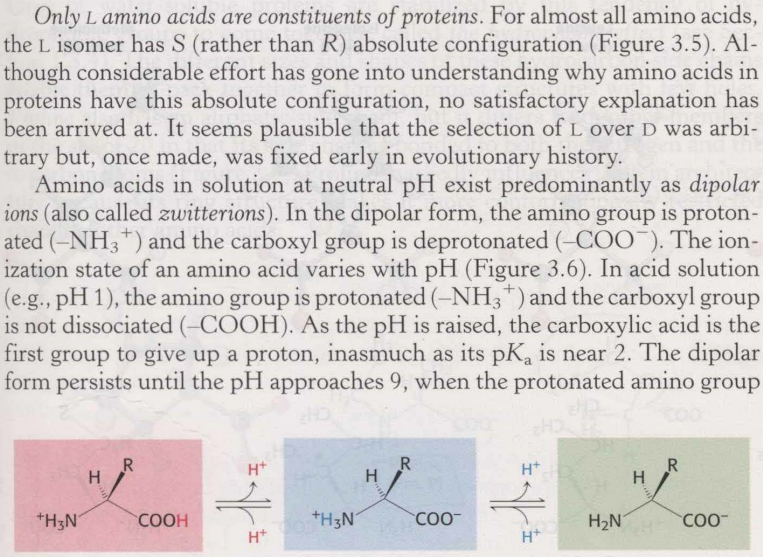


FIGURE 3.6 Ionization state as a function of pH. The ionization state of amino acids is altered by a change in pH. The zwitterionic form predominates near physiological pH.

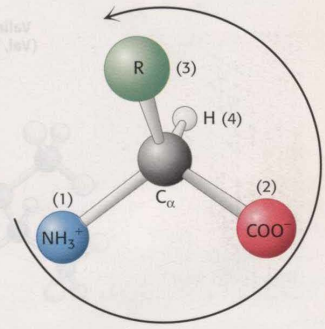


FIGURE 3.5 Only L amino acids are found in proteins. Almost all L amino acids have an *S* absolute configuration (from the Latin *sinister* meaning "left"). The counterclockwise direction of the arrow from highest- to lowest-priority substituents indicates that the chiral center is of the *S* configuration.

Novo Nordisk Ex. 2076, P. 5
Mylan Institutional v. Novo Nordisk
IPR2020-00324

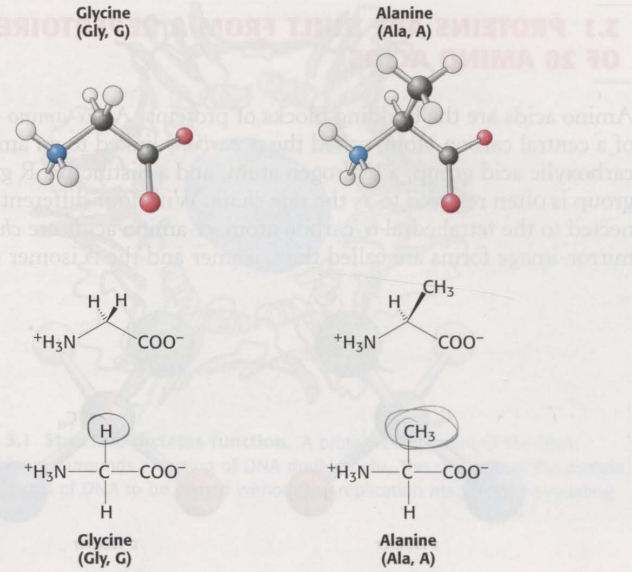


FIGURE 3.7 Structures of glycine and alanine. (Top) Ball-and-stick models show the arrangement of atoms and bonds in space. (Middle) Stereochemically realistic formulas show the geometrical arrangement of bonds around atoms (see Chapter 1 Appendix). (Bottom) Fischer projections show all bonds as being perpendicular for a simplified representation (see Chapter 1 Appendix).

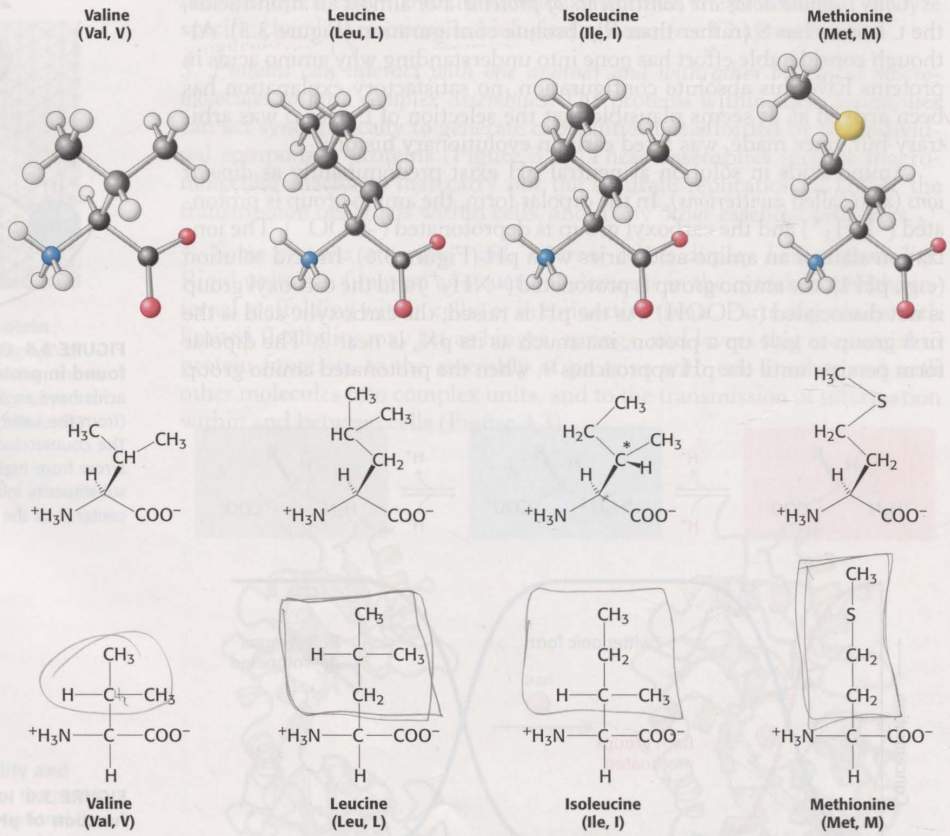


FIGURE 3.8 Amino acids with aliphatic side chains. The additional chiral center of isoleucine is indicated by an asterisk.

Novo Nordisk Ex. 2076, P. 6
Mylan Institutional v. Novo Nordisk
IPR2020-00324

loses a proton. For a review of acid–base concepts and pH, see the appendix to this chapter.

Twenty kinds of side chains varying in *size, shape, charge, hydrogen-bonding capacity, hydrophobic character,* and *chemical reactivity* are commonly found in proteins. Indeed, all proteins in all species—bacterial, archaeal, and eukaryotic—are constructed from the same set of 20 amino acids. This fundamental alphabet of proteins is several billion years old. The remarkable range of functions mediated by proteins results from the diversity and versatility of these 20 building blocks. Understanding how this alphabet is used to create the intricate three-dimensional structures that enable proteins to carry out so many biological processes is an exciting area of biochemistry and one that we will return to in Section 3.6.

Let us look at this set of amino acids. The simplest one is *glycine,* which has just a hydrogen atom as its side chain. With two hydrogen atoms bonded to the α-carbon atom, glycine is unique in being *achiral. Alanine,* the next simplest amino acid, has a methyl group ($-CH_3$) as its side chain (Figure 3.7).

Larger hydrocarbon side chains are found in *valine, leucine,* and *isoleucine* (Figure 3.8). *Methionine* contains a largely *aliphatic* side chain that includes a *thioether* (–S–) group. The side chain of isoleucine includes an additional chiral center; only the isomer shown in Figure 3.8 is found in proteins. The larger aliphatic side chains are *hydrophobic*—that is, they tend to cluster together rather than contact water. The three-dimensional structures of water-soluble proteins are stabilized by this tendency of hydrophobic groups to come together, called *the hydrophobic effect* (see Section 1.3.4). The different sizes and shapes of these hydrocarbon side chains enable them to pack together to form compact structures with few holes. *Proline* also has an aliphatic side chain, but it differs from other members of the set of 20 in that its side chain is bonded to both the nitrogen and the α-carbon atoms (Figure 3.9). Proline markedly influences protein architecture because its ring structure makes it more conformationally restricted than the other amino acids.

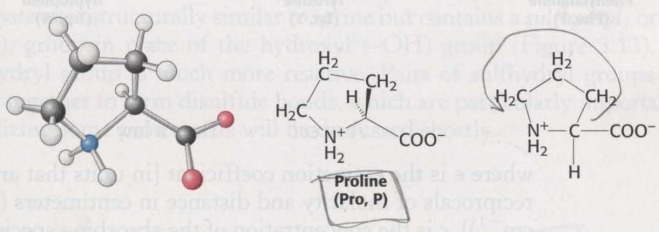


FIGURE 3.9 Cyclic structure of proline. The side chain is joined to both the α carbon and the amino group.

Three amino acids with relatively simple *aromatic side chains* are part of the fundamental repertoire (Figure 3.10). *Phenylalanine,* as its name indicates, contains a phenyl ring attached in place of one of the hydrogens of alanine. The aromatic ring of *tyrosine* contains a hydroxyl group. This hydroxyl group is reactive, in contrast with the rather inert side chains of the other amino acids discussed thus far. *Tryptophan* has an indole ring joined to a methylene ($-CH_2-$) group; the indole group comprises two fused rings and an NH group. Phenylalanine is purely hydrophobic, whereas tyrosine and tryptophan are less so because of their hydroxyl and NH groups. The aromatic rings of tryptophan and tyrosine contain delocalized π electrons that strongly absorb ultraviolet light (Figure 3.11).

A compound's *extinction coefficient* indicates its ability to absorb light. Beer's law gives the absorbance (A) of light at a given wavelength:

Novo Nordisk Ex. 2076, P. 7
Mylan Institutional v. Novo Nordisk
IPR2020-00324

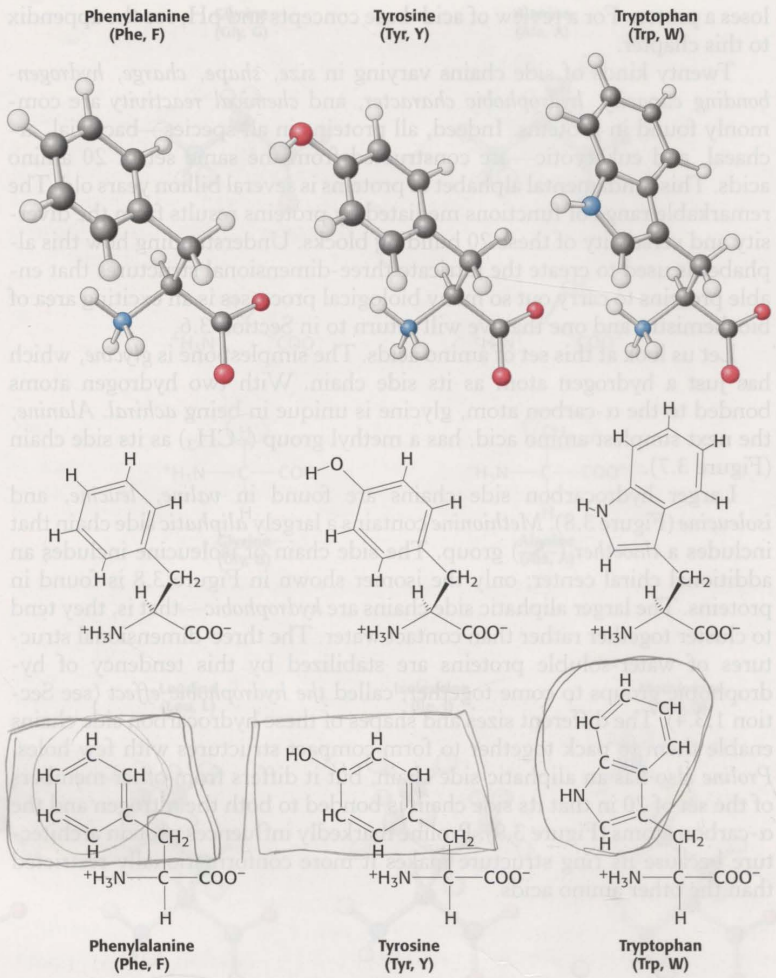


FIGURE 3.10 Amino acids with aromatic side chains. Phenylalanine, tyrosine, and tryptophan have hydrophobic character. Tyrosine and tryptophan also have hydrophilic properties because of their –OH and –NH– groups, respectively.

$$A = \epsilon cl \qquad \text{Beer's law}$$

where ϵ is the extinction coefficient [in units that are the reciprocals of molarity and distance in centimeters (M^{-1} cm^{-1})], c is the concentration of the absorbing species (in units of molarity, M), and l is the length through which the light passes (in units of centimeters). For tryptophan, absorption is maximum at 280 nm and the extinction coefficient is 3400 M^{-1} cm^{-1} whereas, for tyrosine, absorption is maximum at 276 nm and the extinction coefficient is a less-intense 1400 M^{-1} cm^{-1}. Phenylalanine absorbs light less strongly and at shorter wavelengths. The absorption of light at 280 nm can be used to estimate

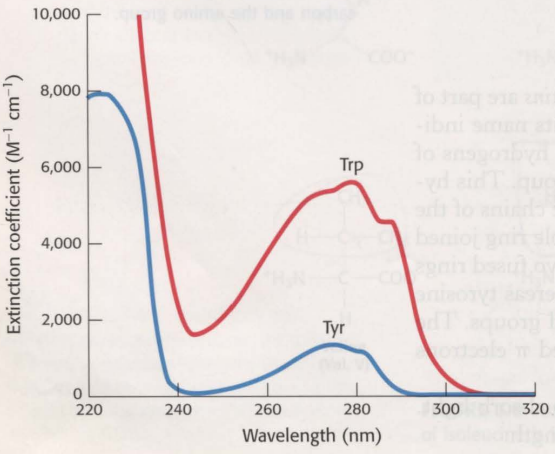


FIGURE 3.11 Absorption spectra of the aromatic amino acids tryptophan (red) and tyrosine (blue). Only these amino acids absorb strongly near 280 nm. [Courtesy of Greg Gatto].

Novo Nordisk Ex. 2076, P. 8
Mylan Institutional v. Novo Nordisk
IPR2020-00324

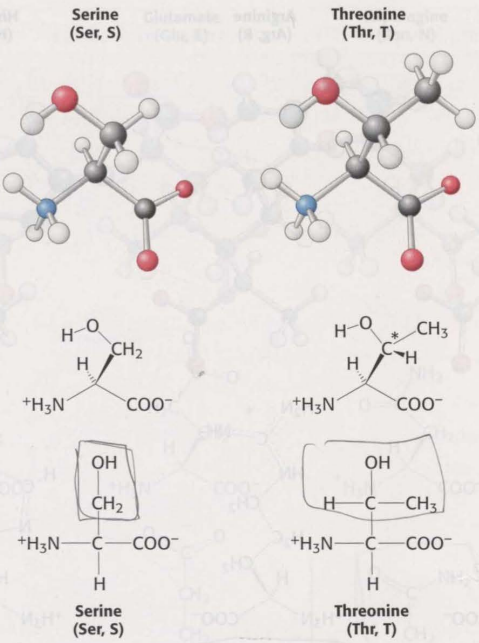


FIGURE 3.12 Amino acids containing aliphatic hydroxyl groups. Serine and threonine contain hydroxyl groups that render them hydrophilic. The additional chiral center in threonine is indicated by an asterisk.

the concentration of a protein in solution if the number of tryptophan and tyrosine residues in the protein is known.

Two amino acids, *serine* and *threonine,* contain aliphatic *hydroxyl groups* (Figure 3.12). Serine can be thought of as a hydroxylated version of alanine, whereas threonine resembles valine with a hydroxyl group in place of one of the valine methyl groups. The hydroxyl groups on serine and threonine make them much more *hydrophilic* (water loving) and *reactive* than alanine and valine. Threonine, like isoleucine, contains an additional asymmetric center; again only one isomer is present in proteins.

Cysteine is structurally similar to serine but contains a *sulfhydryl,* or *thiol* (–SH), group in place of the hydroxyl (–OH) group (Figure 3.13). The sulfhydryl group is much more reactive. Pairs of sulfhydryl groups may come together to form disulfide bonds, which are particularly important in stabilizing some proteins, as will be discussed shortly.

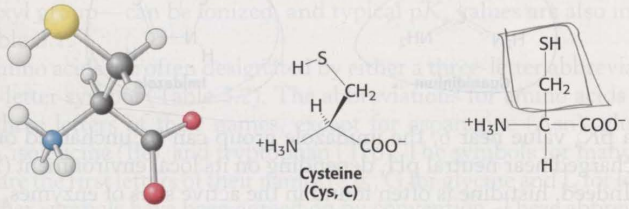


FIGURE 3.13 Structure of cysteine.

We turn now to amino acids with very polar side chains that render them highly hydrophilic. *Lysine* and *arginine* have relatively long side chains that terminate with groups that are *positively charged* at neutral pH. Lysine is capped by a primary amino group and arginine by a guanidinium group. *Histidine* contains an imidazole group, an aromatic ring that also can be positively charged (Figure 3.14).

Novo Nordisk Ex. 2076, P. 9
Mylan Institutional v. Novo Nordisk
IPR2020-00324

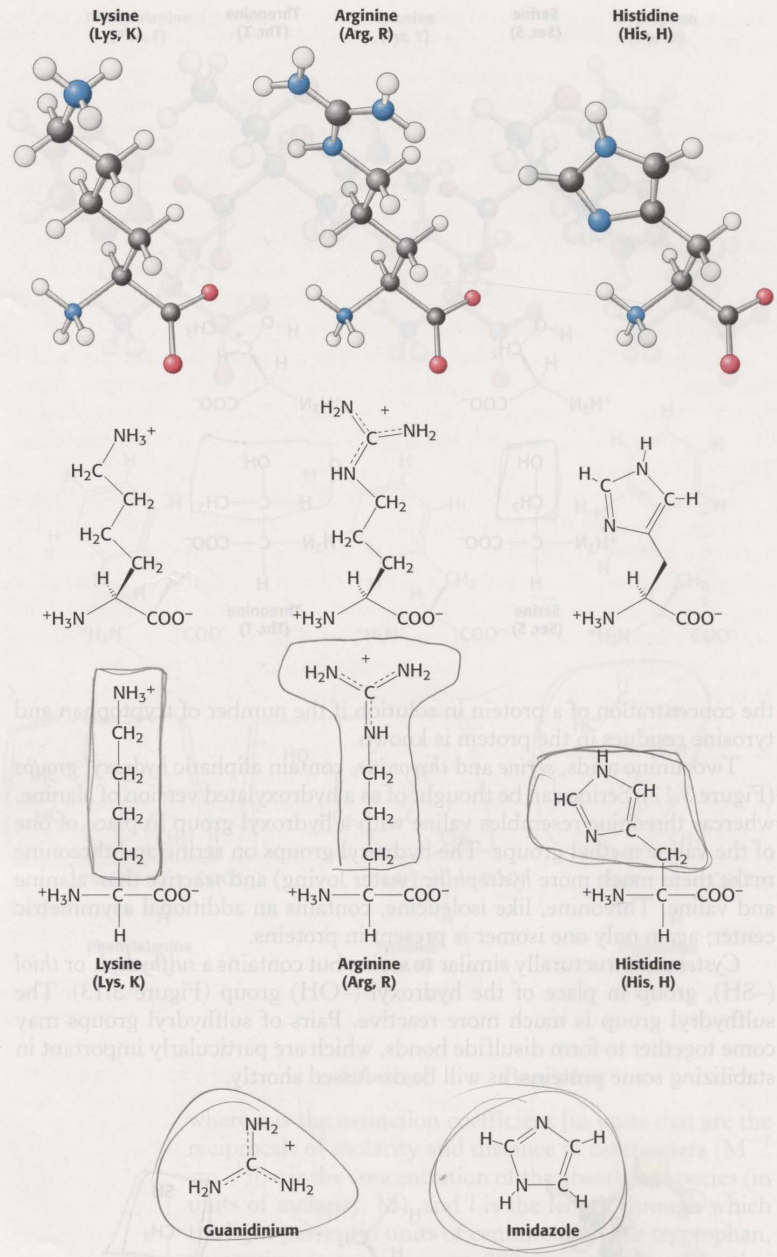


FIGURE 3.14 The basic amino acids lysine, arginine, and histidine.

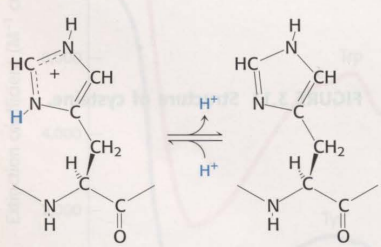


FIGURE 3.15 Histidine ionization.
Histidine can bind or release protons near physiological pH.

With a pK_a value near 6, the imidazole group can be uncharged or positively charged near neutral pH, depending on its local environment (Figure 3.15). Indeed, histidine is often found in the active sites of enzymes, where the imidazole ring can bind and release protons in the course of enzymatic reactions.

The set of amino acids also contains two with *acidic side chains: aspartic acid* and *glutamic acid* (Figure 3.16). These amino acids are often called *aspartate* and *glutamate* to emphasize that their side chains are usually negatively charged at physiological pH. Nonetheless, in some proteins these side chains do accept protons, and this ability is often functionally important. In addition, the set includes uncharged derivatives of aspartate and

Novo Nordisk Ex. 2076, P. 10
Mylan Institutional v. Novo Nordisk
IPR2020-00324

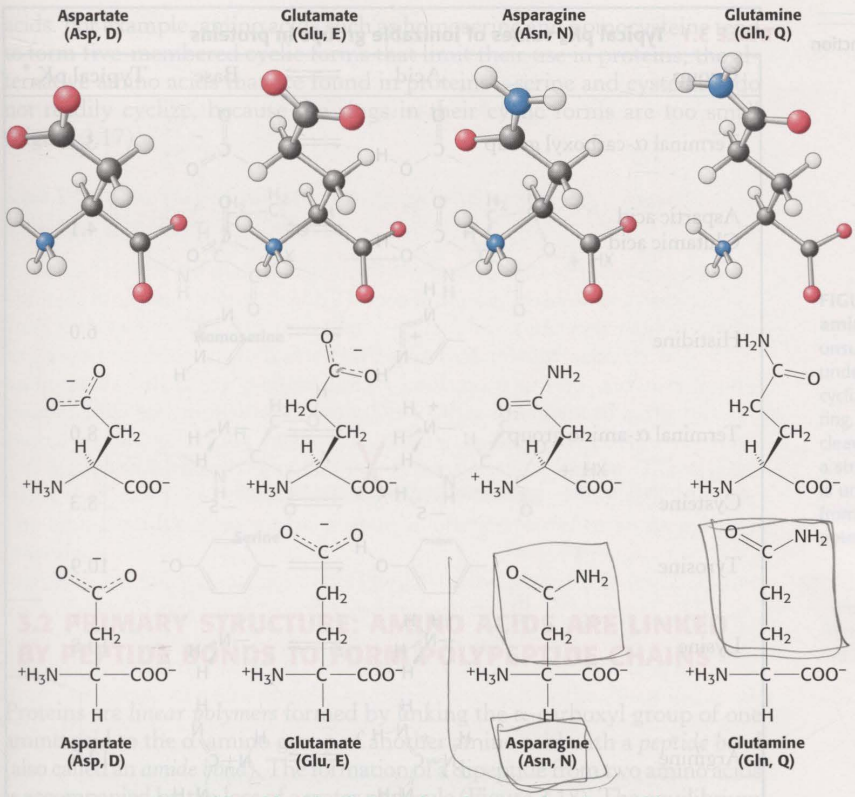


FIGURE 3.16 Amino acids with side-chain carboxylates and carboxamides.

glutamate—*asparagine* and *glutamine*—each of which contains a terminal *carboxamide* in place of a carboxylic acid (Figure 3.16).

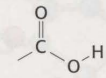
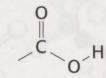
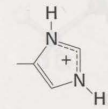
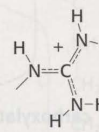
Seven of the 20 amino acids have readily ionizable side chains. These 7 amino acids are able to donate or accept protons to facilitate reactions as well as to form ionic bonds. Table 3.1 gives equilibria and typical pK_a values for ionization of the side chains of tyrosine, cysteine, arginine, lysine, histidine, and aspartic and glutamic acids in proteins. Two other groups in proteins—the terminal α-amino group and the terminal α-carboxyl group—can be ionized, and typical pK_a values are also included in Table 3.1.

Amino acids are often designated by either a three-letter abbreviation or a one-letter symbol (Table 3.2). The abbreviations for amino acids are the first three letters of their names, except for asparagine (Asn), glutamine (Gln), isoleucine (Ile), and tryptophan (Trp). The symbols for many amino acids are the first letters of their names (e.g., G for glycine and L for leucine); the other symbols have been agreed on by convention. These abbreviations and symbols are an integral part of the vocabulary of biochemists.

How did this particular set of amino acids become the building blocks of proteins? First, as a set, they are diverse; their structural and chemical properties span a wide range, endowing proteins with the versatility to assume many functional roles. Second, as noted in Section 2.1.1, many of these amino acids were probably available from prebiotic reactions. Finally, excessive intrinsic reactivity may have eliminated other possible amino

Novo Nordisk Ex. 2076, P. 11
Mylan Institutional v. Novo Nordisk
IPR2020-00324

TABLE 3.1 Typical pK_a values of ionizable groups in proteins

Group	Acid ⇌ Base	Typical pK_a*
Terminal α-carboxyl group		3.1
Aspartic acid Glutamic acid		4.1
Histidine		6.0
Terminal α-amino group		8.0
Cysteine		8.3
Tyrosine		10.9
Lysine		10.8
Arginine		12.5

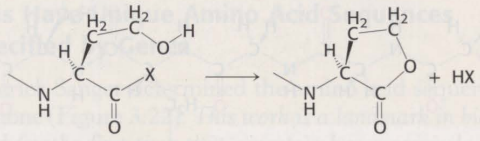
*pK_a values depend on temperature, ionic strength, and the microenvironment of the ionizable group.

TABLE 3.2 Abbreviations for amino acids

Amino acid	Three-letter abbreviation	One-letter abbreviation	Amino acid	Three-letter abbreviation	One-letter abbreviation
Alanine	Ala	A	Methionine	Met	M
Arginine	Arg	R	Phenylalanine	Phe	F
Asparagine	Asn	N	Proline	Pro	P
Aspartic Acid	Asp	D	Serine	Ser	S
Cysteine	Cys	C	Threonine	Thr	T
Glutamine	Gln	Q	Tryptophan	Trp	W
Glutamic Acid	Glu	E	Tyrosine	Tyr	Y
Glycine	Gly	G	Valine	Val	V
Histidine	His	H	Asparagine or aspartic acid	Asx	B
Isoleucine	Ile	I	Glutamine or glutamic acid	Glx	Z
Leucine	Leu	L			
Lysine	Lys	K			

Novo Nordisk Ex. 2076, P. 12
Mylan Institutional v. Novo Nordisk
IPR2020-00324

acids. For example, amino acids such as homoserine and homocysteine tend to form five-membered cyclic forms that limit their use in proteins; the alternative amino acids that are found in proteins—serine and cysteine—do not readily cyclize, because the rings in their cyclic forms are too small (Figure 3.17).



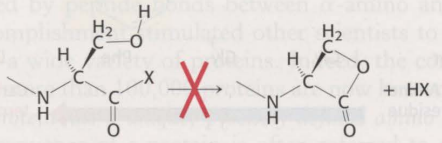


FIGURE 3.17 Undesirable reactivity in amino acids. Some amino acids are unsuitable for proteins because of undesirable cyclization. Homoserine can cyclize to form a stable, five-membered ring, potentially resulting in peptide-bond cleavage. Cyclization of serine would form a strained, four-membered ring and thus is unfavored. X can be an amino group from a neighboring amino acid or another potential leaving group.

3.2 PRIMARY STRUCTURE: AMINO ACIDS ARE LINKED BY PEPTIDE BONDS TO FORM POLYPEPTIDE CHAINS

Proteins are *linear polymers* formed by linking the α-carboxyl group of one amino acid to the α-amino group of another amino acid with a *peptide bond* (also called an *amide bond*). The formation of a dipeptide from two amino acids is accompanied by the loss of a water molecule (Figure 3.18). The equilibrium of this reaction lies on the side of hydrolysis rather than synthesis. Hence, the biosynthesis of peptide bonds requires an input of free energy. Nonetheless, peptide bonds are quite *stable kinetically;* the lifetime of a peptide bond in aqueous solution in the absence of a catalyst approaches 1000 years.

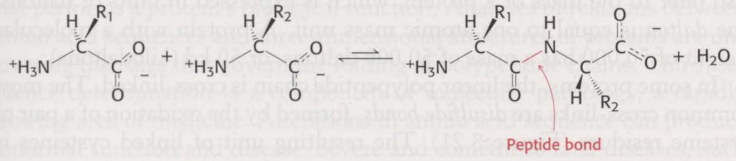


FIGURE 3.18 Peptide-bond formation. The linking of two amino acids is accompanied by the loss of a molecule of water.

A series of amino acids joined by peptide bonds form a *polypeptide chain,* and each amino acid unit in a polypeptide is called a *residue. A polypeptide chain has polarity* because its ends are different, with an α-amino group at one end and an α-carboxyl group at the other. By convention, *the amino end is taken to be the beginning of a polypeptide chain,* and so the sequence of amino acids in a polypeptide chain is written starting with the amino-terminal residue. Thus, in the pentapeptide Tyr-Gly-Gly-Phe-Leu (YGGFL), tyrosine is the amino-terminal (N-terminal) residue and leucine is the carboxyl-terminal (C-terminal) residue (Figure 3.19). Leu-Phe-Gly-Gly-Tyr (LFGGY) is a different pentapeptide, with different chemical properties.

A polypeptide chain consists of a regularly repeating part, called the *main chain* or *backbone,* and a variable part, comprising the distinctive *side chains* (Figure 3.20). The polypeptide backbone is rich in hydrogen-bonding potential. Each residue contains a carbonyl group, which is a good hydrogen-bond acceptor and, with the exception of proline, an NH group, which is a

Novo Nordisk Ex. 2076, P. 13
Mylan Institutional v. Novo Nordisk
IPR2020-00324

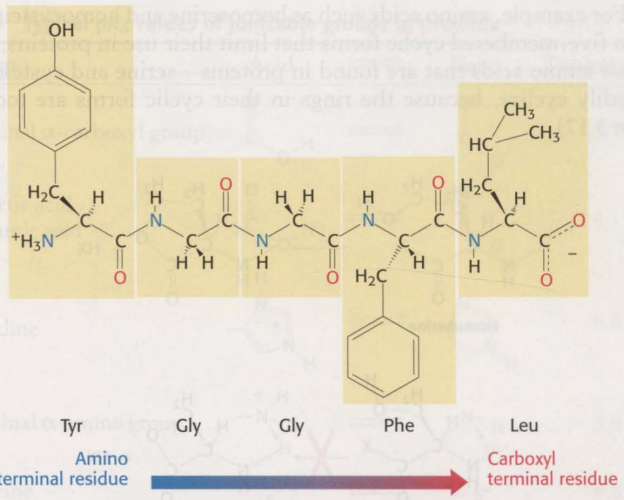


FIGURE 3.19 Amino acid sequences have direction. This illustration of the pentapeptide Try-Gly-Gly-Phe-Leu (YGGFL) shows the sequence from the amino terminus to the carboxyl terminus. This pentapeptide, Leu-enkephalin, is an opioid peptide that modulates the perception of pain. The reverse pentapeptide, Leu-Phe-Gly-Gly-Tyr (LFGGY), is a different molecule and shows no such effects.

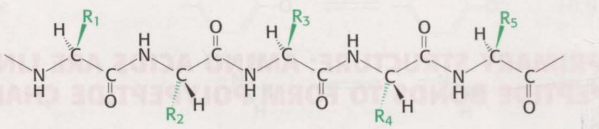


FIGURE 3.20 Components of a polypeptide chain. A polypeptide chain consists of a constant backbone (shown in black) and variable side chains (shown in green).

Dalton–
A unit of mass very nearly equal to that of a hydrogen atom. Named after John Dalton (1766–1844), who developed the atomic theory of matter.

Kilodalton (kd)–
A unit of mass equal to 1000 daltons.

good hydrogen-bond donor. These groups interact with each other and with functional groups from side chains to stabilize particular structures, as will be discussed in detail.

Most natural polypeptide chains contain between 50 and 2000 amino acid residues and are commonly referred to as *proteins*. Peptides made of small numbers of amino acids are called *oligopeptides* or simply *peptides*. The mean molecular weight of an amino acid residue is about 110, and so the molecular weights of most proteins are between 5500 and 220,000. We can also refer to the mass of a protein, which is expressed in units of daltons; one *dalton* is equal to one atomic mass unit. A protein with a molecular weight of 50,000 has a mass of 50,000 daltons, or 50 kd (kilodaltons).

In some proteins, the linear polypeptide chain is cross-linked. The most common cross-links are *disulfide bonds*, formed by the oxidation of a pair of cysteine residues (Figure 3.21). The resulting unit of linked cysteines is

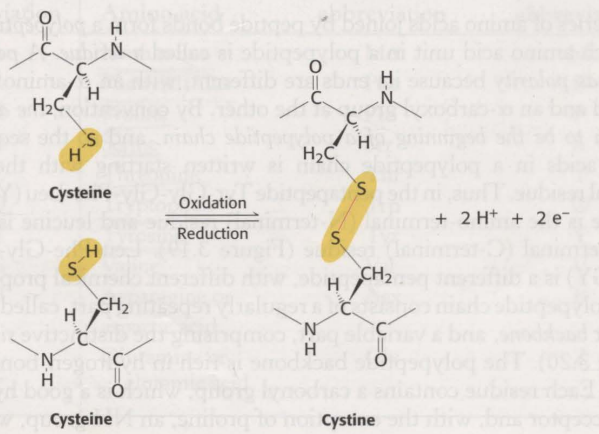


FIGURE 3.21 Cross-links. The formation of a disulfide bond from two cysteine residues is an oxidation reaction.

Novo Nordisk Ex. 2076, P. 14
Mylan Institutional v. Novo Nordisk
IPR2020-00324

called *cystine*. Extracellular proteins often have several disulfide bonds, whereas intracellular proteins usually lack them. Rarely, nondisulfide cross-links derived from other side chains are present in some proteins. For example, collagen fibers in connective tissue are strengthened in this way, as are fibrin blood clots.

3.2.1 Proteins Have Unique Amino Acid Sequences That Are Specified by Genes

In 1953, Frederick Sanger determined the amino acid sequence of insulin, a protein hormone (Figure 3.22). *This work is a landmark in biochemistry because it showed for the first time that a protein has a precisely defined amino acid sequence.* Moreover, it demonstrated that insulin consists only of L amino acids linked by peptide bonds between α-amino and α-carboxyl groups. This accomplishment stimulated other scientists to carry out sequence studies of a wide variety of proteins. Indeed, the complete amino acid sequences of more than 100,000 proteins are now known. *The striking fact is that each protein has a unique, precisely defined amino acid sequence.* The amino acid sequence of a protein is often referred to as its *primary structure*.

A series of incisive studies in the late 1950s and early 1960s revealed that the amino acid sequences of proteins are genetically determined. The sequence of nucleotides in DNA, the molecule of heredity, specifies a complementary sequence of nucleotides in RNA, which in turn specifies the amino acid sequence of a protein. In particular, each of the 20 amino acids of the repertoire is encoded by one or more specific sequences of three nucleotides (Section 5.5).

Knowing amino acid sequences is important for several reasons. First, knowledge of the sequence of a protein is usually essential to elucidating its mechanism of action (e.g., the catalytic mechanism of an enzyme). Moreover, proteins with novel properties can be generated by varying the sequence of known proteins. Second, amino acid sequences determine the three-dimensional structures of proteins. Amino acid sequence is the link between the genetic message in DNA and the three-dimensional structure that performs a protein's biological function. Analyses of relations between amino acid sequences and three-dimensional structures of proteins are uncovering the rules that govern the folding of polypeptide chains. Third, sequence determination is a component of molecular pathology, a rapidly growing area of medicine. Alterations in amino acid sequence can produce abnormal function and disease. Severe and sometimes fatal diseases, such as sickle-cell anemia and cystic fibrosis, can result from a change in a single amino acid within a protein. Fourth, the sequence of a protein reveals much about its evolutionary history (see Chapter 7). Proteins resemble one another in amino acid sequence only if they have a common ancestor. Consequently, molecular events in evolution can be traced from amino acid sequences; molecular paleontology is a flourishing area of research.

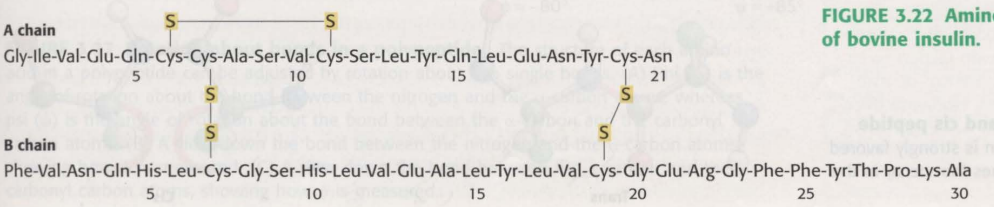


FIGURE 3.22 Amino acid sequence of bovine insulin.

Novo Nordisk Ex. 2076, P. 15
Mylan Institutional v. Novo Nordisk
IPR2020-00324

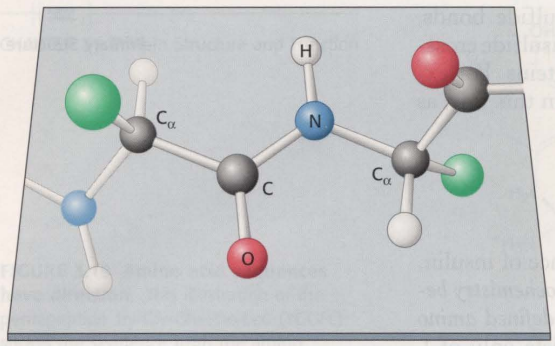
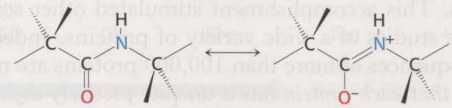


FIGURE 3.23 Peptide bonds are planar. In a pair of linked amino acids, six atoms (C_α, C, O, N, H, and C_α) lie in a plane. Side chains are shown as green balls.

3.2.2 Polypeptide Chains Are Flexible Yet Conformationally Restricted

Examination of the geometry of the protein backbone reveals several important features. First, *the peptide bond is essentially planar* (Figure 3.23). Thus, for a pair of amino acids linked by a peptide bond, six atoms lie in the same plane: the α-carbon atom and CO group from the first amino acid and the NH group and α-carbon atom from the second amino acid. The nature of the chemical bonding within a peptide explains this geometric preference. The peptide bond has considerable *double-bond character*, which prevents rotation about this bond.



Peptide bond resonance structures

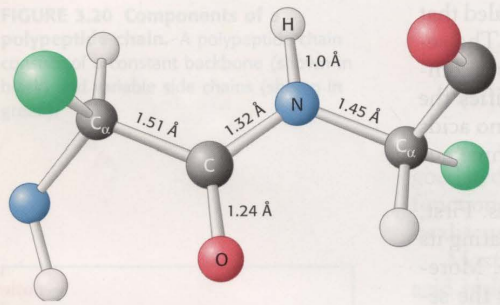
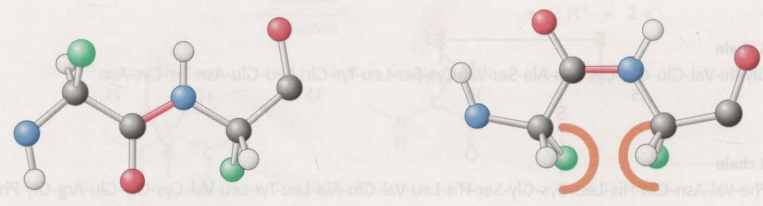


FIGURE 3.24 Typical bond lengths within a peptide unit. The peptide unit is shown in the trans configuration.

The inability of the bond to rotate constrains the conformation of the peptide backbone and accounts for the bond's planarity. This double-bond character is also expressed in the length of the bond between the CO and NH groups. The C–N distance in a peptide bond is typically 1.32 Å, which is between the values expected for a C–N single bond (1.49 Å) and a C=N double bond (1.27 Å), as shown in Figure 3.24. Finally, the peptide bond is uncharged, allowing polymers of amino acids linked by peptide bonds to form tightly packed globular structures.

Two configurations are possible for a planar peptide bond. In the trans configuration, the two α-carbon atoms are on opposite sides of the peptide bond. In the cis configuration, these groups are on the same side of the peptide bond. *Almost all peptide bonds in proteins are trans*. This preference for trans over cis can be explained by the fact that steric clashes between groups attached to the α-carbon atoms hinder formation of the cis form but do not occur in the trans configuration (Figure 3.25). By far the most common cis peptide bonds are X–Pro linkages. Such bonds show less preference for the trans configuration because the nitrogen of proline is bonded to two tetrahedral carbon atoms, limiting the steric differences between the trans and cis forms (Figure 3.26).

In contrast with the peptide bond, the bonds between the amino group and the α-carbon atom and between the α-carbon atom and the carbonyl group are pure single bonds. The two adjacent rigid peptide units may rotate about these bonds, taking on various orientations. *This freedom of rotation*



Trans Cis

FIGURE 3.25 Trans and cis peptide bonds. The trans form is strongly favored because of steric clashes that occur in the cis form.

Novo Nordisk Ex. 2076, P. 16
Mylan Institutional v. Novo Nordisk
IPR2020-00324

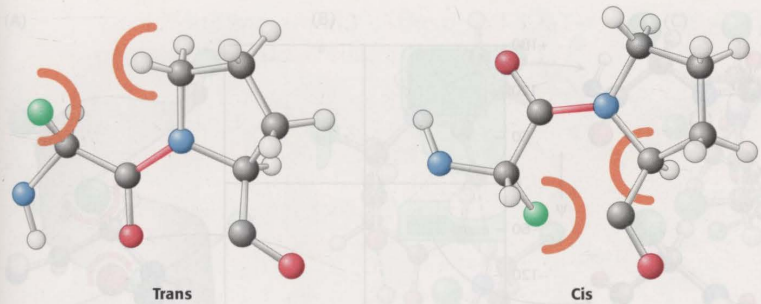


FIGURE 3.26 Trans and cis X–Pro bonds. The energies of these forms are relatively balanced because steric clashes occur in both forms.

about two bonds of each amino acid allows proteins to fold in many different ways. The rotations about these bonds can be specified by dihedral angles (Figure 3.27). The angle of rotation about the bond between the nitrogen and the α-carbon atoms is called *phi* (ϕ). The angle of rotation about the bond between the α-carbon and the carbonyl carbon atoms is called *psi* (ψ). A clockwise rotation about either bond as viewed from the front of the back group corresponds to a positive value. The ϕ and ψ angles determine the path of the polypeptide chain.

Dihedral angle–
A measure of the rotation about a bond, usually taken to lie between −180° and +180°. Dihedral angles are sometimes called torsion angles.

Are all combinations of ϕ and ψ possible? G. N. Ramachandran recognized that many combinations are forbidden because of steric collisions between atoms. The allowed values can be visualized on a two-dimensional plot called a *Ramachandran diagram* (Figure 3.28). Three-quarters of the possible (ϕ, ψ) combinations are excluded simply by local steric clashes. *Steric exclusion, the fact that two atoms cannot be in the same place at the same time, can be a powerful organizing principle.*

The ability of biological polymers such as proteins to fold into well-defined structures is remarkable thermodynamically. Consider the equilibrium between an unfolded polymer that exists as a random coil—that is, as a mixture of many possible conformations—and the folded form that adopts a unique conformation. The favorable entropy associated with the large number of conformations in the unfolded form opposes folding and must be overcome by interactions favoring the folded form. Thus, highly flexible polymers with a large number of possible conformations do not fold into unique structures. *The rigidity of the peptide unit and the restricted set of allowed ϕ and ψ angles limits the number of structures accessible to the unfolded form sufficiently to allow protein folding to occur.*

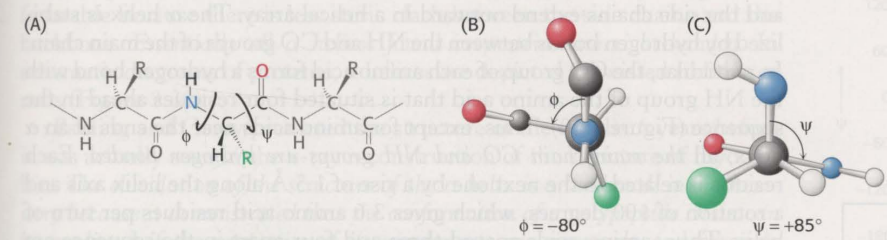


FIGURE 3.27 Rotation about bonds in a polypeptide. The structure of each amino acid in a polypeptide can be adjusted by rotation about two single bonds. (A) Phi (ϕ) is the angle of rotation about the bond between the nitrogen and the α-carbon atoms, whereas psi (ψ) is the angle of rotation about the bond between the α-carbon and the carbonyl carbon atoms. (B) A view down the bond between the nitrogen and the α-carbon atoms, showing how ϕ is measured. (C) A view down the bond between the α-carbon and the carbonyl carbon atoms, showing how ψ is measured.

Novo Nordisk Ex. 2076, P. 17
Mylan Institutional v. Novo Nordisk
IPR2020-00324

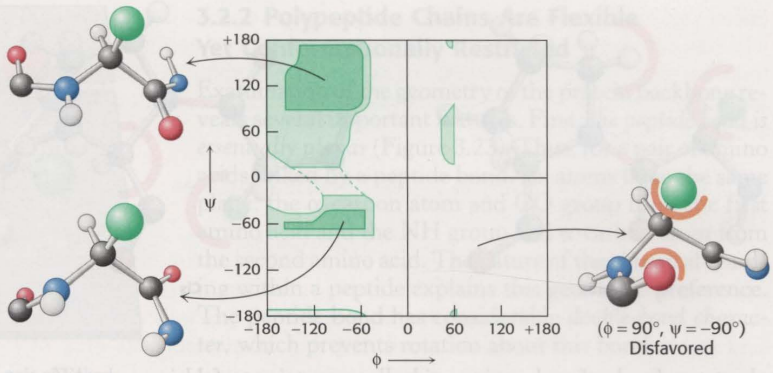


FIGURE 3.28 A Ramachandran diagram showing the values of ϕ and ψ. Not all ϕ and ψ values are possible without collisions between atoms. The most favorable regions are shown in dark green; borderline regions are shown in light green. The structure on the right is disfavored because of steric clashes.

3.3 SECONDARY STRUCTURE: POLYPEPTIDE CHAINS CAN FOLD INTO REGULAR STRUCTURES SUCH AS THE ALPHA HELIX, THE BETA SHEET, AND TURNS AND LOOPS

Can a polypeptide chain fold into a regularly repeating structure? In 1951, Linus Pauling and Robert Corey proposed two periodic structures called the α *helix* (alpha helix) and the β *pleated sheet* (beta pleated sheet). Subsequently, other structures such as the β *turn* and *omega* (Ω) *loop* were identified. Although not periodic, these common turn or loop structures are well defined and contribute with α helices and β sheets to form the final protein structure.

STRUCTURAL INSIGHTS, appearing throughout the book, are molecular modeling-based tutorials that enable you to review structure and learn what the latest research tells us about the workings of the molecule. To access, go to the Web site: www.whfreeman.com/biochem5, and select the chapter, Structural Insights, and the title.

STRUCTURAL INSIGHTS, Elements of Protein Structure *provides interactive representations of some of the important elements of protein architecture described in this chapter, including a summary of secondary structure motifs.*

3.3.1 The Alpha Helix Is a Coiled Structure Stabilized by Intrachain Hydrogen Bonds

In evaluating potential structures, Pauling and Corey considered which conformations of peptides were sterically allowed and which most fully exploited the hydrogen-bonding capacity of the backbone NH and CO groups. The first of their proposed structures, the α *helix*, is a rodlike structure (Figure 3.29). A tightly coiled backbone forms the inner part of the rod and the side chains extend outward in a helical array. The α helix is stabilized by hydrogen bonds between the NH and CO groups of the main chain. In particular, the CO group of each amino acid forms a hydrogen bond with the NH group of the amino acid that is situated four residues ahead in the sequence (Figure 3.30). Thus, except for amino acids near the ends of an α helix, all *the main-chain CO and NH groups are hydrogen bonded*. Each residue is related to the next one by a rise of 1.5 Å along the helix axis and a rotation of 100 degrees, which gives 3.6 amino acid residues per turn of helix. Thus, amino acids spaced three and four apart in the sequence are spatially quite close to one another in an α helix. In contrast, amino acids two apart in the sequence are situated on opposite sides of the helix and so are unlikely to make contact. The *pitch* of the α helix, which is equal to the product of the translation (1.5 Å) and the number of residues per turn (3.6), is 5.4 Å. The *screw sense* of a helix can be right-handed (clockwise) or left-handed (counterclockwise). The Ramachandran diagram reveals that both

Screw sense—
Describes the direction in which a helical structure rotates with respect to its axis. If, viewed down the axis of a helix, the chain turns in a clockwise direction, it has a right-handed screw sense. If the turning is counterclockwise, the screw sense is left-handed.

Novo Nordisk Ex. 2076, P. 18
Mylan Institutional v. Novo Nordisk
IPR2020-00324

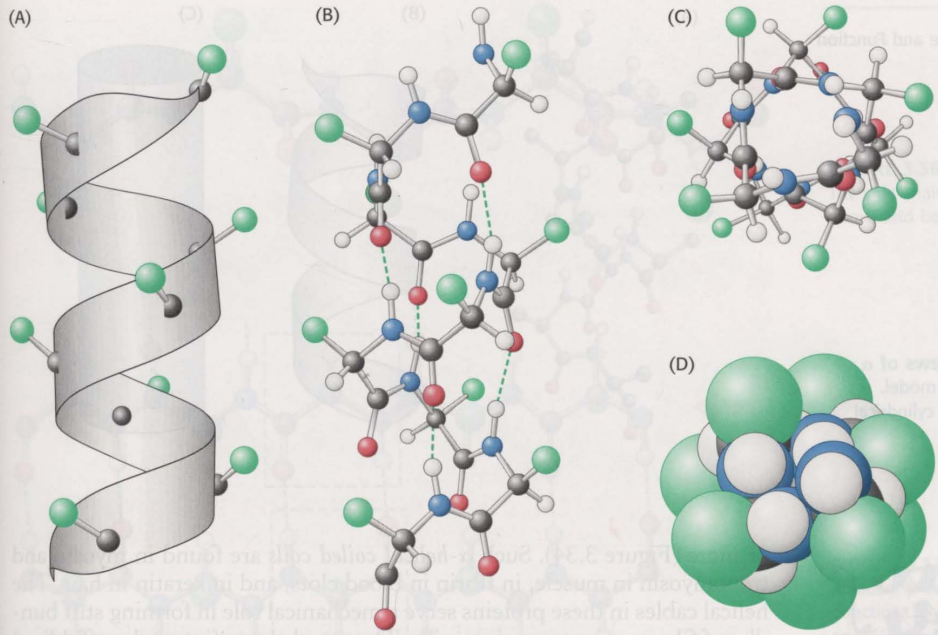


FIGURE 3.29 Structure of the α helix. (A) A ribbon depiction with the α-carbon atoms and side chains (green) shown. (B) A side view of a ball-and-stick version depicts the hydrogen bonds (dashed lines) between NH and CO groups. (C) An end view shows the coiled backbone as the inside of the helix and the side chains (green) projecting outward. (D) A space-filling view of part C shows the tightly packed interior core of the helix.

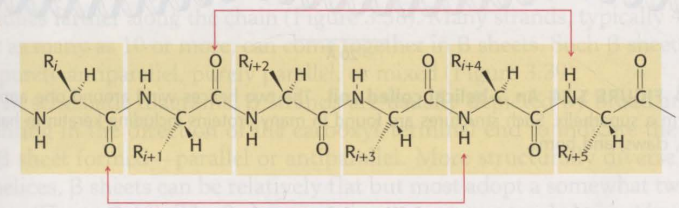


FIGURE 3.30 Hydrogen-bonding scheme for an α helix. In the α helix, the CO group of residue *n* forms a hydrogen bond with the NH group of residue *n* + 4.

the right-handed and the left-handed helices are among allowed conformations (Figure 3.31). However, right-handed helices are energetically more favorable because there is less steric clash between the side chains and the backbone. *Essentially all α helices found in proteins are right-handed.* In schematic diagrams of proteins, α helices are depicted as twisted ribbons or rods (Figure 3.32).

Pauling and Corey predicted the structure of the α helix 6 years before it was actually seen in the x-ray reconstruction of the structure of myoglobin. *The elucidation of the structure of the α helix is a landmark in biochemistry because it demonstrated that the conformation of a polypeptide chain can be predicted if the properties of its components are rigorously and precisely known.*

The α-helical content of proteins ranges widely, from nearly none to almost 100%. For example, about 75% of the residues in ferritin, a protein that helps store iron, are in α helices (Figure 3.33). Single α helices are usually less than 45 Å long. However, two or more α helices can entwine to form a very stable structure, which can have a length of 1000 Å (100 nm, or 0.1 μm)

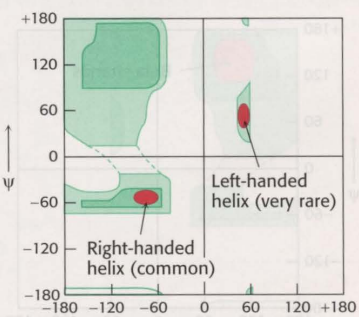


FIGURE 3.31 Ramachandran diagram for helices. Both right- and left-handed helices lie in regions of allowed conformations in the Ramachandran diagram. However, essentially all α helices in proteins are right-handed.

Novo Nordisk Ex. 2076, P. 19
Mylan Institutional v. Novo Nordisk
IPR2020-00324

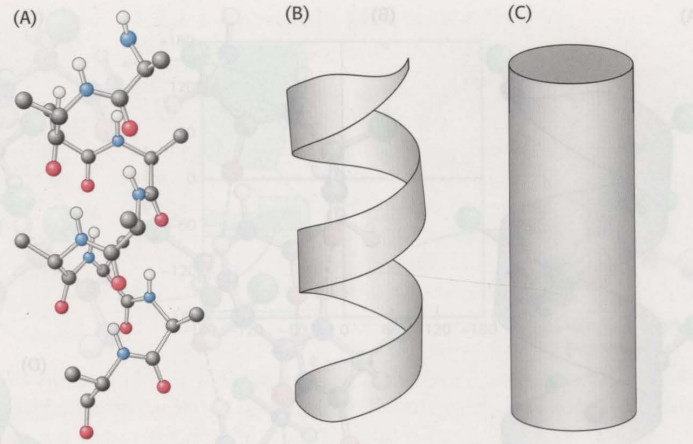


FIGURE 3.32 Schematic views of α helices. (A) A ball-and-stick model. (B) A ribbon depiction. (C) A cylindrical depiction.

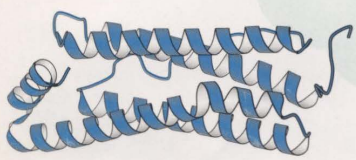


FIGURE 3.33 A largely α helical protein. Ferritin, an iron-storage protein, is built from a bundle of α helices.

or more (Figure 3.34). Such *α-helical coiled coils* are found in myosin and tropomyosin in muscle, in fibrin in blood clots, and in keratin in hair. The helical cables in these proteins serve a mechanical role in forming stiff bundles of fibers, as in porcupine quills. The cytoskeleton (internal scaffolding) of cells is rich in so-called intermediate filaments, which also are two-stranded α-helical coiled coils. Many proteins that span biological membranes also contain α helices.

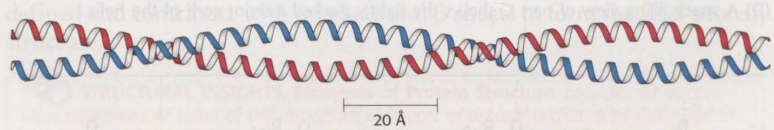


FIGURE 3.34 An α-helical coiled coil. The two helices wind around one another to form a superhelix. Such structures are found in many proteins including keratin in hair, quills, claws, and horns.

3.3.2 Beta Sheets Are Stabilized by Hydrogen Bonding Between Polypeptide Strands

Pauling and Corey discovered another periodic structural motif, which they named the *β pleated sheet* (β because it was the second structure that they elucidated, the α helix having been the first). The β pleated sheet (or, more simply, the β sheet) differs markedly from the rodlike α helix. A polypeptide chain, called a *β strand*, in a β sheet is almost fully extended rather than being tightly coiled as in the α helix. A range of extended structures are sterically allowed (Figure 3.35).

The distance between adjacent amino acids along a β strand is approximately 3.5 Å, in contrast with a distance of 1.5 Å along an α helix. The side chains of adjacent amino acids point in opposite directions (Figure 3.36). A β sheet is formed by linking two or more β strands by hydrogen bonds. Adjacent chains in a β sheet can run in opposite directions (antiparallel β sheet) or in the same direction (parallel β sheet). In the antiparallel arrangement, the NH group and the CO group of each amino acid are respectively hydrogen bonded to the CO group and the NH group of a partner on the adjacent chain (Figure 3.37). In the parallel arrangement, the hydrogen-bonding scheme is slightly more complicated. For each amino acid, the NH group is hydrogen

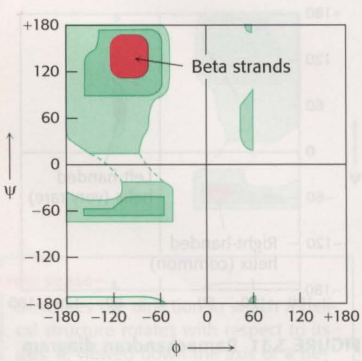


FIGURE 3.35 Ramachandran diagram for β strands. The red area shows the sterically allowed conformations of extended, β-strand-like structures.

Novo Nordisk Ex. 2076, P. 20
Mylan Institutional v. Novo Nordisk
IPR2020-00324

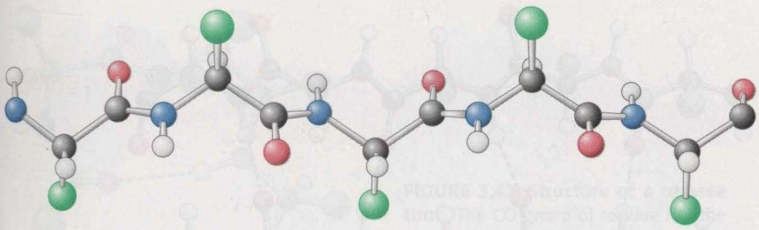


FIGURE 3.36 Structure of a β strand. The side chains (green) are alternately above and below the plane of the strand.

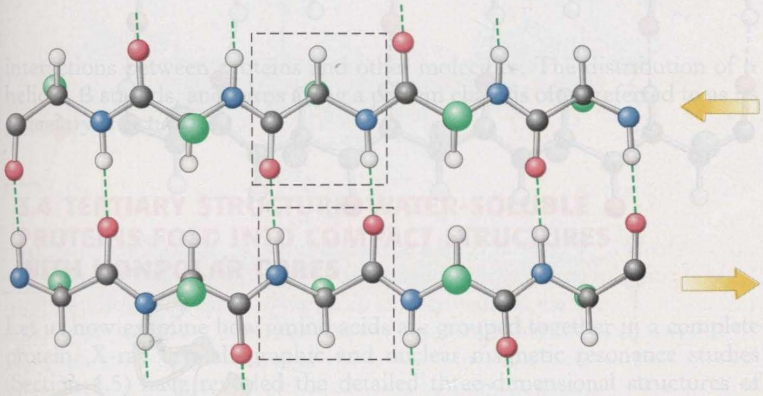


FIGURE 3.37 An antiparallel β sheet. Adjacent β strands run in opposite directions. Hydrogen bonds between NH and CO groups connect each amino acid to a single amino acid on an adjacent strand, stabilizing the structure.

bonded to the CO group of one amino acid on the adjacent strand, whereas the CO group is hydrogen bonded to the NH group on the amino acid two residues farther along the chain (Figure 3.38). Many strands, typically 4 or 5 but as many as 10 or more, can come together in β sheets. Such β sheets can be purely antiparallel, purely parallel, or mixed (Figure 3.39).

In schematic diagrams, β strands are usually depicted by broad arrows pointing in the direction of the carboxyl-terminal end to indicate the type of β sheet formed—parallel or antiparallel. More structurally diverse than α helices, β sheets can be relatively flat but most adopt a somewhat twisted shape (Figure 3.40). The β sheet is an important structural element in many proteins. For example, fatty acid-binding proteins, important for lipid metabolism, are built almost entirely from β sheets (Figure 3.41).

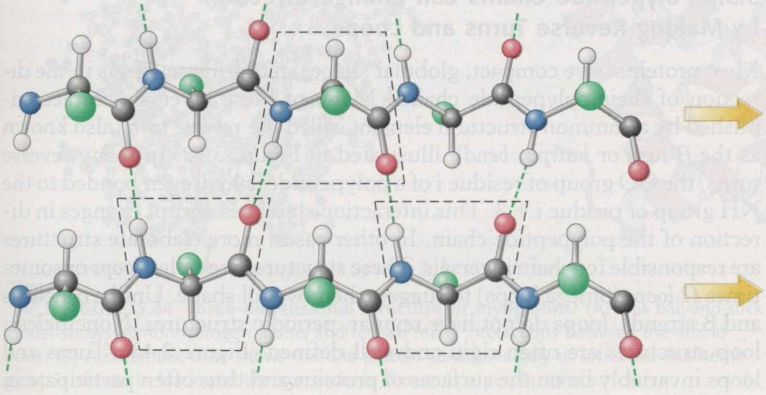


FIGURE 3.38 A parallel β sheet. Adjacent β strands run in the same direction. Hydrogen bonds connect each amino acid on one strand with two different amino acids on the adjacent strand.

Novo Nordisk Ex. 2076, P. 21
Mylan Institutional v. Novo Nordisk
IPR2020-00324

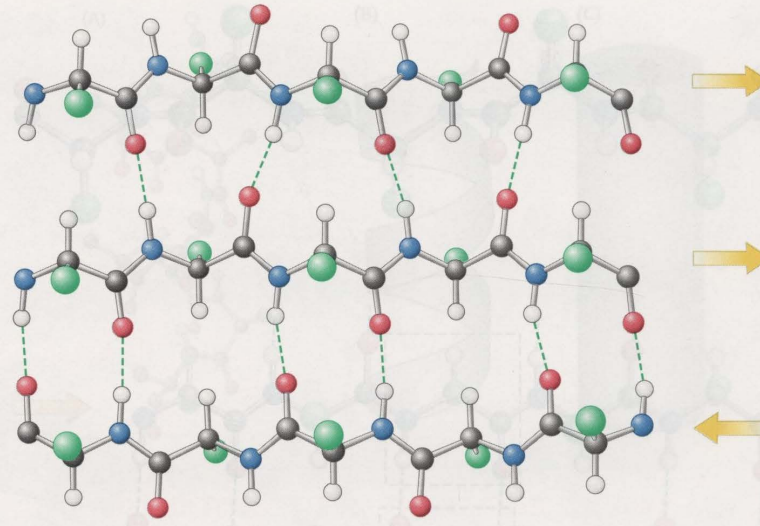


FIGURE 3.39 Structure of a mixed β sheet.

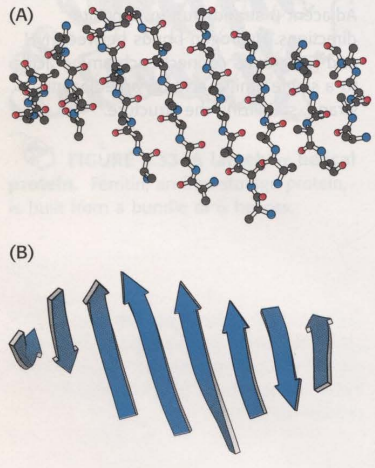


FIGURE 3.40 A twisted β sheet. (A) A ball-and-stick model. (B) A schematic model. (C) The schematic view rotated by 90 degrees to illustrate the twist more clearly.

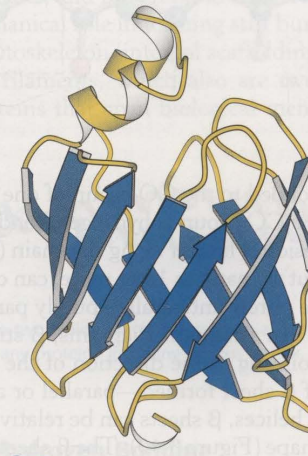


FIGURE 3.41 A protein rich in β sheets. The structure of a fatty acid-binding protein.

3.3.3 Polypeptide Chains Can Change Direction by Making Reverse Turns and Loops

Most proteins have compact, globular shapes, requiring reversals in the direction of their polypeptide chains. Many of these reversals are accomplished by a common structural element called the *reverse turn* (also known as the *β turn* or *hairpin bend*), illustrated in Figure 3.42. In many reverse turns, the CO group of residue i of a polypeptide is hydrogen bonded to the NH group of residue $i + 3$. This interaction stabilizes abrupt changes in direction of the polypeptide chain. In other cases, more elaborate structures are responsible for chain reversals. These structures are called *loops* or sometimes *Ω loops* (omega loops) to suggest their overall shape. Unlike α helices and β strands, loops do not have regular, periodic structures. Nonetheless, loop structures are often rigid and well defined (Figure 3.43). Turns and loops invariably lie on the surfaces of proteins and thus often participate in

Novo Nordisk Ex. 2076, P. 22
Mylan Institutional v. Novo Nordisk
IPR2020-00324

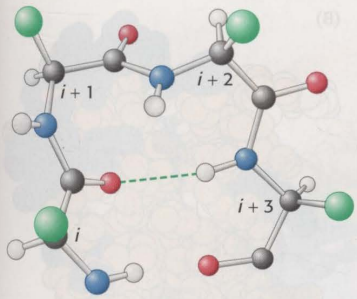


FIGURE 3.42 Structure of a reverse turn. The CO group of residue *i* of the polypeptide chain is hydrogen bonded to the NH group of residue *i* + 3 to stabilize the turn.

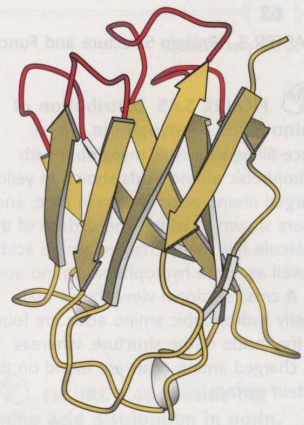


FIGURE 3.43 Loops on a protein surface. A part of an antibody molecule has surface loops (shown in red) that mediate interactions with other molecules.

interactions between proteins and other molecules. The distribution of α helices, β strands, and turns along a protein chain is often referred to as its *secondary structure*.

3.4 TERTIARY STRUCTURE: WATER-SOLUBLE PROTEINS FOLD INTO COMPACT STRUCTURES WITH NONPOLAR CORES

Let us now examine how amino acids are grouped together in a complete protein. X-ray crystallographic and nuclear magnetic resonance studies (Section 4.5) have revealed the detailed three-dimensional structures of thousands of proteins. We begin here with a preview of *myoglobin*, the first protein to be seen in atomic detail.

Myoglobin, the oxygen carrier in muscle, is a single polypeptide chain of 153 amino acids (see also Chapters 7 and 10). The capacity of myoglobin to bind oxygen depends on the presence of *heme*, a nonpolypeptide *prosthetic (helper) group* consisting of protoporphyrin IX and a central iron atom. *Myoglobin is an extremely compact molecule*. Its overall dimensions are 45 × 35 × 25 Å, an order of magnitude less than if it were fully stretched out (Figure 3.44).

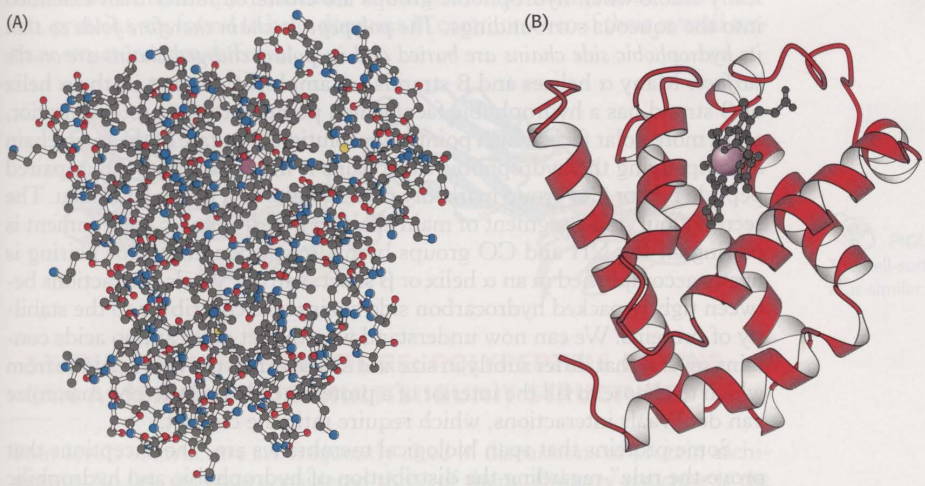


FIGURE 3.44 Three-dimensional structure of myoglobin. (A) This ball-and-stick model shows all nonhydrogen atoms and reveals many interactions between the amino acids. (B) A schematic view shows that the protein consists largely of α helices. The heme group is shown in black and the iron atom is shown as a purple sphere.

Novo Nordisk Ex. 2076, P. 23
Mylan Institutional v. Novo Nordisk
IPR2020-00324

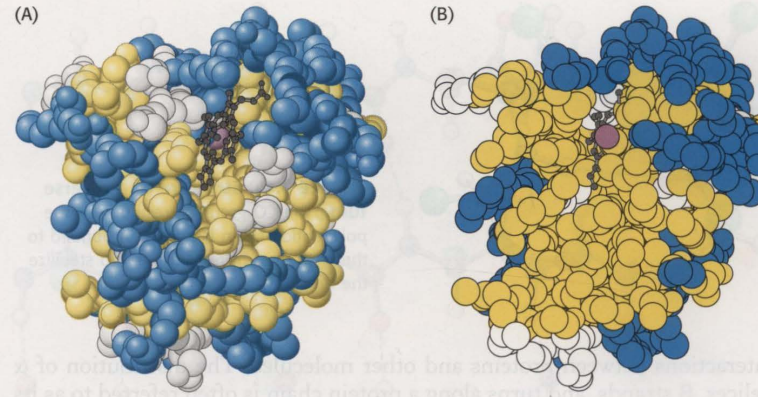


FIGURE 3.45 Distribution of amino acids in myoglobin. (A) A space-filling model of myoglobin with hydrophobic amino acids shown in yellow, charged amino acids shown in blue, and others shown in white. The surface of the molecule has many charged amino acids, as well as some hydrophobic amino acids. (B) A cross-sectional view shows that mostly hydrophobic amino acids are found on the inside of the structure, whereas the charged amino acids are found on the protein surface.

About 70% of the main chain is folded into eight α helices, and much of the rest of the chain forms turns and loops between helices.

The folding of the main chain of myoglobin, like that of most other proteins, is complex and devoid of symmetry. The overall course of the polypeptide chain of a protein is referred to as its *tertiary structure*. A unifying principle emerges from the distribution of side chains. The striking fact is that *the interior consists almost entirely of nonpolar residues* such as leucine, valine, methionine, and phenylalanine (Figure 3.45). Charged residues such as aspartate, glutamate, lysine, and arginine are absent from the inside of myoglobin. The only polar residues inside are two histidine residues, which play critical roles in binding iron and oxygen. The outside of myoglobin, on the other hand, consists of both polar and nonpolar residues. The space-filling model shows that there is very little empty space inside.

This contrasting distribution of polar and nonpolar residues reveals a key facet of protein architecture. In an aqueous environment, protein folding is driven by the strong tendency of hydrophobic residues to be excluded from water (see Section 1.3.4). Recall that a system is more thermodynamically stable when hydrophobic groups are clustered rather than extended into the aqueous surroundings. *The polypeptide chain therefore folds so that its hydrophobic side chains are buried and its polar, charged chains are on the surface.* Many α helices and β strands are amphipathic; that is, the α helix or β strand has a hydrophobic face, which points into the protein interior, and a more polar face, which points into solution. The fate of the main chain accompanying the hydrophobic side chains is important, too. An unpaired peptide NH or CO group markedly prefers water to a nonpolar milieu. The secret of burying a segment of main chain in a hydrophobic environment is pairing all the NH and CO groups by hydrogen bonding. This pairing is neatly accomplished in an α helix or β sheet. Van der Waals interactions between tightly packed hydrocarbon side chains also contribute to the stability of proteins. We can now understand why the set of 20 amino acids contains several that differ subtly in size and shape. They provide a palette from which to choose to fill the interior of a protein neatly and thereby maximize van der Waals interactions, which require intimate contact.

Some proteins that span biological membranes are "the exceptions that prove the rule" regarding the distribution of hydrophobic and hydrophilic amino acids throughout three-dimensional structures. For example, consider porins, proteins found in the outer membranes of many bacteria (Figure 3.46). The permeability barriers of membranes are built largely of alkane chains that are quite hydrophobic (Section 12.4). Thus, porins are

Novo Nordisk Ex. 2076, P. 24
Mylan Institutional v. Novo Nordisk
IPR2020-00324

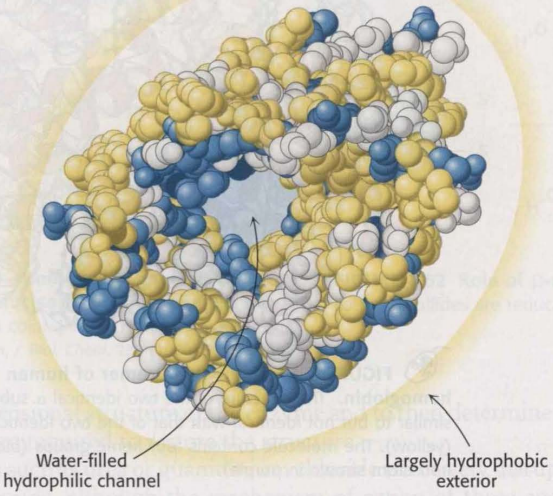


FIGURE 3.46 "Inside out" amino acid distribution in porin. The outside of porin (which contacts hydrophobic groups in membranes) is covered largely with hydrophobic residues, whereas the center includes a water-filled channel lined with charged and polar amino acids.

covered on the outside largely with hydrophobic residues that interact with the neighboring alkane chains. In contrast, the center of the protein contains many charged and polar amino acids that surround a water-filled channel running through the middle of the protein. Thus, because porins function in hydrophobic environments, they are "inside out" relative to proteins that function in aqueous solution.

Some polypeptide chains fold into two or more compact regions that may be connected by a flexible segment of polypeptide chain, rather like pearls on a string. These compact globular units, called *domains*, range in size from about 30 to 400 amino acid residues. For example, the extracellular part of CD4, the cell-surface protein on certain cells of the immune system to which the human immunodeficiency virus (HIV) attaches itself, comprises four similar domains of approximately 100 amino acids each (Figure 3.47). Often, proteins are found to have domains in common even if their overall tertiary structures are different.

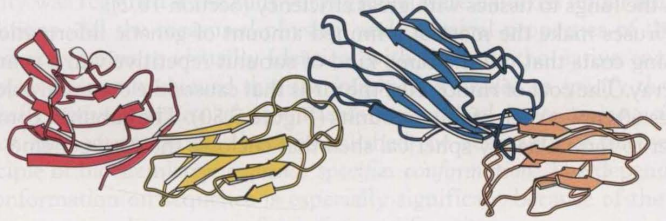


FIGURE 3.47 Protein domains. The cell-surface protein CD4 consists of four similar domains.

3.5 QUATERNARY STRUCTURE: POLYPEPTIDE CHAINS CAN ASSEMBLE INTO MULTISUBUNIT STRUCTURES

Four levels of structure are frequently cited in discussions of protein architecture. So far, we have considered three of them. *Primary structure* is the amino acid sequence. *Secondary structure* refers to the spatial arrangement of amino acid residues that are nearby in the sequence. Some of these arrangements are of a regular kind, giving rise to a periodic structure. The α helix and β strand are elements of secondary structure. *Tertiary structure*

Novo Nordisk Ex. 2076, P. 25
Mylan Institutional v. Novo Nordisk
IPR2020-00324

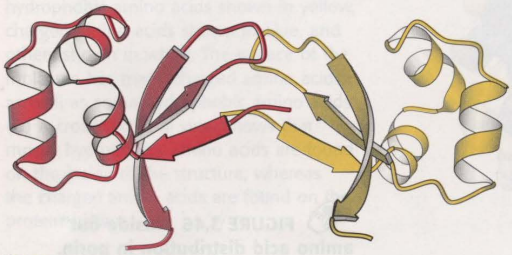


FIGURE 3.48 Quaternary structure. The Cro protein of bacteriophage λ is a dimer of identical subunits.

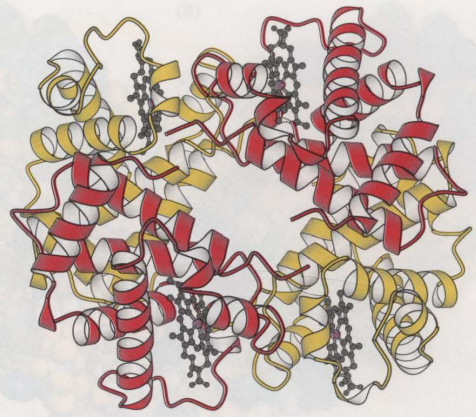


FIGURE 3.49 The $\alpha_2\beta_2$ tetramer of human hemoglobin. The structure of the two identical α subunits (red) is similar to but not identical with that of the two identical β subunits (yellow). The molecule contains four heme groups (black with the iron atom shown in purple).

refers to the spatial arrangement of amino acid residues that are far apart in the sequence and to the pattern of disulfide bonds. We now turn to proteins containing more than one polypeptide chain. Such proteins exhibit a fourth level of structural organization. Each polypeptide chain in such a protein is called a *subunit*. *Quaternary structure* refers to the spatial arrangement of subunits and the nature of their interactions. The simplest sort of quaternary structure is a *dimer*, consisting of two identical subunits. This organization is present in the DNA-binding protein Cro found in a bacterial virus called λ (Figure 3.48). More complicated quaternary structures also are common. More than one type of subunit can be present, often in variable numbers. For example, human hemoglobin, the oxygen-carrying protein in blood, consists of two subunits of one type (designated α) and two subunits of another type (designated β), as illustrated in Figure 3.49. Thus, the hemoglobin molecule exists as an $\alpha_2\beta_2$ tetramer. Subtle changes in the arrangement of subunits within the hemoglobin molecule allow it to carry oxygen from the lungs to tissues with great efficiency (Section 10.2).

Viruses make the most of a limited amount of genetic information by forming coats that use the same kind of subunit repetitively in a symmetric array. The coat of rhinovirus, the virus that causes the common cold, includes 60 copies each of four subunits (Figure 3.50). The subunits come together to form a nearly spherical shell that encloses the viral genome.

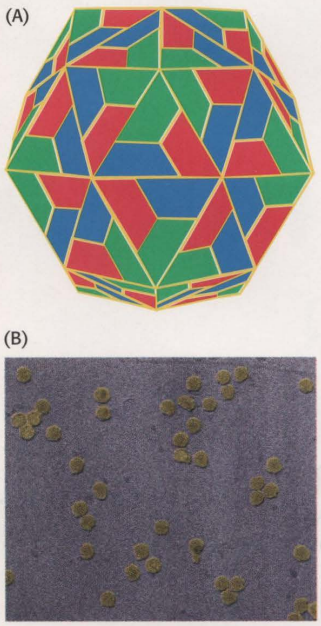


FIGURE 3.50 Complex quaternary structure. The coat of rhinovirus comprises 60 copies of each of four subunits. (A) A schematic view depicting the three types of subunits (shown in red, blue, and green) visible from outside the virus. (B) An electron micrograph showing rhinovirus particles. [Courtesy of Norm Olson, Dept. of Biological Sciences, Purdue University.]

3.6 THE AMINO ACID SEQUENCE OF A PROTEIN DETERMINES ITS THREE-DIMENSIONAL STRUCTURE

How is the elaborate three-dimensional structure of proteins attained, and how is the three-dimensional structure related to the one-dimensional amino acid sequence information? The classic work of Christian Anfinsen in the 1950s on the enzyme ribonuclease revealed the relation between the amino acid sequence of a protein and its conformation. Ribonuclease is a single polypeptide chain consisting of 124 amino acid residues cross-linked by four disulfide bonds (Figure 3.51). Anfinsen's plan was to destroy the

Novo Nordisk Ex. 2076, P. 26
Mylan Institutional v. Novo Nordisk
IPR2020-00324

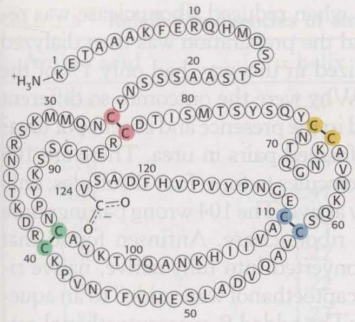


FIGURE 3.51 Amino acid sequence of bovine ribonuclease. The four disulfide bonds are shown in color. [After C. H. W. Hirs, S. Moore, and W. H. Stein, *J. Biol. Chem.* 235 (1960):633.]

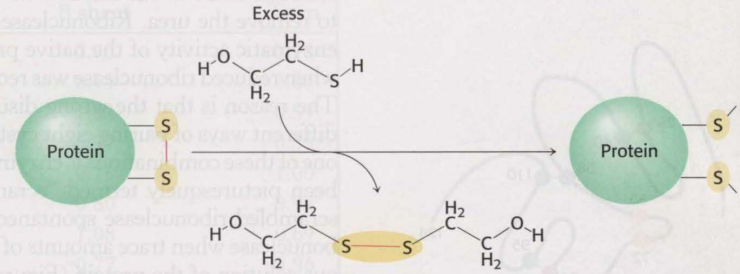


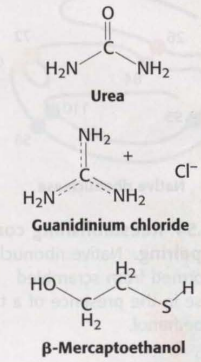
FIGURE 3.52 Role of β-mercaptoethanol in reducing disulfide bonds. Note that, as the disulfides are reduced, the β-mercaptoethanol is oxidized and forms dimers.

three-dimensional structure of the enzyme and to then determine what conditions were required to restore the structure.

Agents such as urea or guanidinium chloride effectively disrupt the noncovalent bonds, although the mechanism of action of these agents is not fully understood. The disulfide bonds can be cleaved reversibly by reducing them with a reagent such as *β-mercaptoethanol* (Figure 3.52). In the presence of a large excess of β-mercaptoethanol, a protein is produced in which the disulfides (cystines) are fully converted into sulfhydryls (cysteines).

Most polypeptide chains devoid of cross-links assume a *random-coil conformation* in 8 M urea or 6 M guanidinium chloride, as evidenced by physical properties such as viscosity and optical activity. When ribonuclease was treated with β-mercaptoethanol in 8 M urea, the product was a fully reduced, randomly coiled polypeptide chain *devoid of enzymatic activity*. In other words, ribonuclease was *denatured* by this treatment (Figure 3.53).

Anfinsen then made the critical observation that the denatured ribonuclease, freed of urea and β-mercaptoethanol by dialysis, slowly regained enzymatic activity. He immediately perceived the significance of this chance finding: the sulfhydryl groups of the denatured enzyme became oxidized by air, and the enzyme spontaneously refolded into a catalytically active form. Detailed studies then showed that nearly all the original enzymatic activity was regained if the sulfhydryl groups were oxidized under suitable conditions. All the measured physical and chemical properties of the refolded enzyme were virtually identical with those of the native enzyme. These experiments showed that *the information needed to specify the catalytically active structure of ribonuclease is contained in its amino acid sequence*. Subsequent studies have established the generality of this central principle of biochemistry: *sequence specifies conformation*. The dependence of conformation on sequence is especially significant because of the intimate connection between conformation and function.



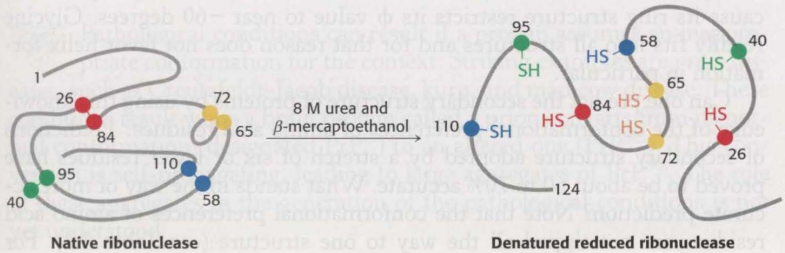


FIGURE 3.53 Reduction and denaturation of ribonuclease.

Novo Nordisk Ex. 2076, P. 27
Mylan Institutional v. Novo Nordisk
IPR2020-00324

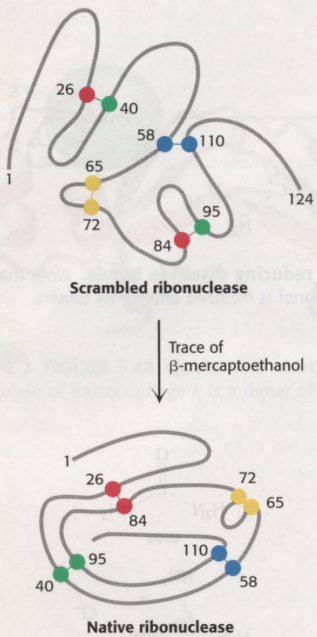


FIGURE 3.54 Reestablishing correct disulfide pairing. Native ribonuclease can be reformed from scrambled ribonuclease in the presence of a trace of β-mercaptoethanol.

A quite different result was obtained when reduced ribonuclease was reoxidized while it was still in 8 M urea and the preparation was then dialyzed to remove the urea. Ribonuclease reoxidized in this way had only 1% of the enzymatic activity of the native protein. Why were the outcomes so different when reduced ribonuclease was reoxidized in the presence and absence of urea? The reason is that the wrong disulfides formed pairs in urea. There are 105 different ways of pairing eight cysteine molecules to form four disulfides; only one of these combinations is enzymatically active. The 104 wrong pairings have been picturesquely termed "scrambled" ribonuclease. Anfinsen found that scrambled ribonuclease spontaneously converted into fully active, native ribonuclease when trace amounts of β-mercaptoethanol were added to an aqueous solution of the protein (Figure 3.54). The added β-mercaptoethanol catalyzed the rearrangement of disulfide pairings until the native structure was regained in about 10 hours. *This process was driven by the decrease in free energy as the scrambled conformations were converted into the stable, native conformation of the enzyme.* The native disulfide pairings of ribonuclease thus contribute to the stabilization of the thermodynamically preferred structure.

Similar refolding experiments have been performed on many other proteins. In many cases, the native structure can be generated under suitable conditions. For other proteins, however, refolding does not proceed efficiently. In these cases, the unfolding protein molecules usually become tangled up with one another to form aggregates. Inside cells, proteins called *chaperones* block such illicit interactions (Section 11.3.6).

3.6.1 Amino Acids Have Different Propensities for Forming Alpha Helices, Beta Sheets, and Beta Turns

How does the amino acid sequence of a protein specify its three-dimensional structure? How does an unfolded polypeptide chain acquire the form of the native protein? These fundamental questions in biochemistry can be approached by first asking a simpler one: What determines whether a particular sequence in a protein forms an α helix, a β strand, or a turn? Examining the frequency of occurrence of particular amino acid residues in these secondary structures (Table 3.3) can be a source of insight into this determination. Residues such as alanine, glutamate, and leucine tend to be present in α helices, whereas valine and isoleucine tend to be present in β strands. Glycine, asparagine, and proline have a propensity for being in turns.

The results of studies of proteins and synthetic peptides have revealed some reasons for these preferences. The α helix can be regarded as the default conformation. Branching at the β-carbon atom, as in valine, threonine, and isoleucine, tends to destabilize α helices because of steric clashes. These residues are readily accommodated in β strands, in which their side chains project out of the plane containing the main chain. Serine, aspartate, and asparagine tend to disrupt α helices because their side chains contain hydrogen-bond donors or acceptors in close proximity to the main chain, where they compete for main-chain NH and CO groups. Proline tends to disrupt both α helices and β strands because it lacks an NH group and because its ring structure restricts its ϕ value to near −60 degrees. Glycine readily fits into all structures and for that reason does not favor helix formation in particular.

Can one predict the secondary structure of proteins by using this knowledge of the conformational preferences of amino acid residues? Predictions of secondary structure adopted by a stretch of six or fewer residues have proved to be about 60 to 70% accurate. What stands in the way of more accurate prediction? Note that the conformational preferences of amino acid residues are not tipped all the way to one structure (see Table 3.3). For

Novo Nordisk Ex. 2076, P. 28
Mylan Institutional v. Novo Nordisk
IPR2020-00324

TABLE 3.3 Relative frequencies of amino acid residues in secondary structures

Amino acid	α helix	β sheet	Turn
Ala	**1.29**	0.90	0.78
Cys	**1.11**	0.74	0.80
Leu	**1.30**	1.02	0.59
Met	**1.47**	0.97	0.39
Glu	**1.44**	0.75	1.00
Gln	**1.27**	0.80	0.97
His	**1.22**	1.08	0.69
Lys	**1.23**	0.77	0.96
Val	0.91	**1.49**	0.47
Ile	0.97	**1.45**	0.51
Phe	1.07	**1.32**	0.58
Tyr	0.72	**1.25**	1.05
Trp	0.99	**1.14**	0.75
Thr	0.82	**1.21**	1.03
Gly	0.56	0.92	**1.64**
Ser	0.82	0.95	**1.33**
Asp	1.04	0.72	**1.41**
Asn	0.90	0.76	**1.28**
Pro	0.52	0.64	**1.91**
Arg	0.96	0.99	0.88

Note: The amino acids are grouped according to their preference for α helices (top group), β sheets (second group), or turns (third group). Arginine shows no significant preference for any of the structures.
After T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2d ed. (W. H. Freeman and Company, 1992), p. 256.

example, glutamate, one of the strongest helix formers, prefers α helix to β strand by only a factor of two. The preference ratios of most other residues are smaller. Indeed, some penta- and hexapeptide sequences have been found to adopt one structure in one protein and an entirely different structure in another (Figure 3.55). Hence, some amino acid sequences do not uniquely determine secondary structure. Tertiary interactions—interactions between residues that are far apart in the sequence—may be decisive in specifying the secondary structure of some segments. *The context is often crucial in determining the conformational outcome.* The conformation of a protein evolved to work in a particular environment or context.

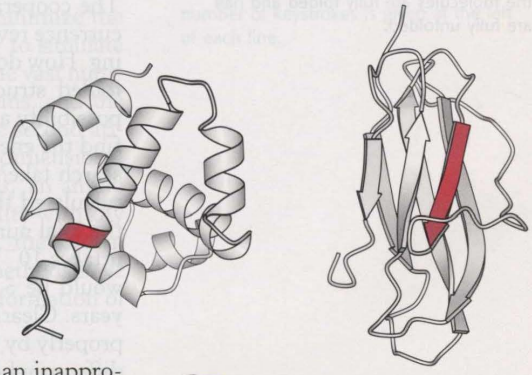


FIGURE 3.55 Alternative conformations of a peptide sequence. Many sequences can adopt alternative conformations in different proteins. Here the sequence VDLLKN shown in red assumes an α helix in one protein context (left) and a β strand in another (right).

Pathological conditions can result if a protein assumes an inappropriate conformation for the context. Striking examples are *prion diseases,* such as Creutzfeldt-Jacob disease, kuru, and mad cow disease. These conditions result when a brain protein called a prion converts from its normal conformation (designated PrP^{C}) to an altered one (PrP^{SC}). This conversion is self-propagating, leading to large aggregates of PrP^{SC}. The role of these aggregates in the generation of the pathological conditions is not yet understood.

Novo Nordisk Ex. 2076, P. 29
Mylan Institutional v. Novo Nordisk
IPR2020-00324

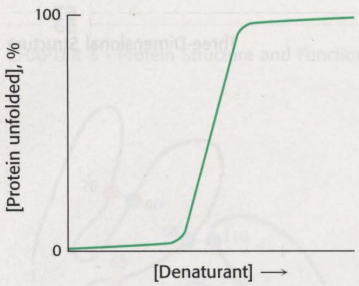


FIGURE 3.56 Transition from folded to unfolded state. Most proteins show a sharp transition from the folded to unfolded form on treatment with increasing concentrations of denaturants.

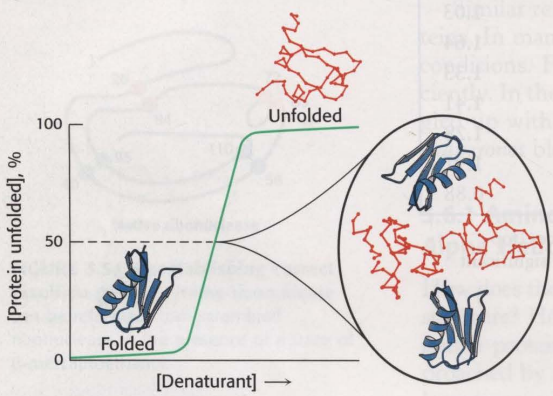


FIGURE 3.57 Components of a partially denatured protein solution. In a half-unfolded protein solution, half the molecules are fully folded and half are fully unfolded.

3.6.2 Protein Folding Is a Highly Cooperative Process

As stated earlier, proteins can be denatured by heat or by chemical denaturants such as urea or guanidium chloride. For many proteins, a comparison of the degree of unfolding as the concentration of denaturant increases has revealed a relatively sharp transition from the folded, or native, form to the unfolded, or denatured, form, suggesting that only these two conformational states are present to any significant extent (Figure 3.56). A similar sharp transition is observed if one starts with unfolded proteins and removes the denaturants, allowing the proteins to fold.

Protein folding and unfolding is thus largely an *"all or none" process* that results from a *cooperative transition*. For example, suppose that a protein is placed in conditions under which some part of the protein structure is thermodynamically unstable. As this part of the folded structure is disrupted, the interactions between it and the remainder of the protein will be lost. The loss of these interactions, in turn, will destabilize the remainder of the structure. Thus, conditions that lead to the disruption of any part of a protein structure are likely to unravel the protein completely. The structural properties of proteins provide a clear rationale for the cooperative transition.

The consequences of cooperative folding can be illustrated by considering the contents of a protein solution under conditions corresponding to the middle of the transition between the folded and unfolded forms. Under these conditions, the protein is "half folded." Yet the solution will contain no half-folded molecules but, instead, will be a 50/50 mixture of fully folded and fully unfolded molecules (Figure 3.57). Structures that are partly intact and partly disrupted are not thermodynamically stable and exist only transiently. Cooperative folding ensures that partly folded structures that might interfere with processes within cells do not accumulate.

3.6.3 Proteins Fold by Progressive Stabilization of Intermediates Rather Than by Random Search

The cooperative folding of proteins is a thermodynamic property; its occurrence reveals nothing about the kinetics and mechanism of protein folding. How does a protein make the transition from a diverse ensemble of unfolded structures into a unique conformation in the native form? One possibility a priori would be that all possible conformations are tried out to find the energetically most favorable one. How long would such a random search take? Consider a small protein with 100 residues. Cyrus Levinthal calculated that, if each residue can assume three different conformations, the total number of structures would be 3^{100}, which is equal to 5×10^{47}. If it takes 10^{-13} s to convert one structure into another, the total search time would be $5 \times 10^{47} \times 10^{-13}$ s, which is equal to 5×10^{34} s, or 1.6×10^{27} years. Clearly, it would take much too long for even a small protein to fold properly by randomly trying out all possible conformations. The enormous difference between calculated and actual folding times is called *Levinthal's paradox*.

The way out of this dilemma is to recognize the power of *cumulative selection*. Richard Dawkins, in *The Blind Watchmaker*, asked how long it would take a monkey poking randomly at a typewriter to reproduce Hamlet's remark to Polonius, "Methinks it is like a weasel" (Figure 3.58). An astronomically large number of keystrokes, of the order of 10^{40}, would be required. However, suppose that we preserved each correct character and

Novo Nordisk Ex. 2076, P. 30
Mylan Institutional v. Novo Nordisk
IPR2020-00324

allowed the monkey to retype only the wrong ones. In this case, only a few thousand keystrokes, on average, would be needed. The crucial difference between these cases is that the first employs a completely random search, whereas, in the second, *partly correct intermediates are retained.*

The essence of protein folding is the retention of partly correct intermediates. However, the protein-folding problem is much more difficult than the one presented to our simian Shakespeare. First, the criterion of correctness is not a residue-by-residue scrutiny of conformation by an omniscient observer but rather the total free energy of the transient species. Second, proteins are only marginally stable. The free-energy difference between the folded and the unfolded states of a typical 100-residue protein is 10 kcal mol^{-1} (42 kJ mol^{-1}), and thus each residue contributes on average only 0.1 kcal mol^{-1} (0.42 kJ mol^{-1}) of energy to maintain the folded state. This amount is less than that of thermal energy, which is 0.6 kcal mol^{-1} (2.5 kJ mol^{-1}) at room temperature. This meager stabilization energy means that correct intermediates, especially those formed early in folding, can be lost. The analogy is that the monkey would be somewhat free to undo its correct keystrokes. Nonetheless, the interactions that lead to cooperative folding can stabilize intermediates as structure builds up. Thus, local regions, which have significant structural preference, though not necessarily stable on their own, will tend to adopt their favored structures and, as they form, can interact with one other, leading to increasing stabilization.

3.6.4 Prediction of Three-Dimensional Structure from Sequence Remains a Great Challenge

The amino acid sequence completely determines the three-dimensional structure of a protein. However, the prediction of three-dimensional structure from sequence has proved to be extremely difficult. As we have seen, the local sequence appears to determine only between 60% and 70% of the secondary structure; long-range interactions are required to fix the full secondary structure and the tertiary structure.

Investigators are exploring two fundamentally different approaches to predicting three-dimensional structure from amino acid sequence. The first is *ab initio prediction,* which attempts to predict the folding of an amino acid sequence without any direct reference to other known protein structures. Computer-based calculations are employed that attempt to minimize the free energy of a structure with a given amino acid sequence or to simulate the folding process. The utility of these methods is limited by the vast number of possible conformations, the marginal stability of proteins, and the subtle energetics of weak interactions in aqueous solution. The second approach takes advantage of our growing knowledge of the three-dimensional structures of many proteins. In these *knowledge-based methods,* an amino acid sequence of unknown structure is examined for compatibility with any known protein structures. If a significant match is detected, the known structure can be used as an initial model. Knowledge-based methods have been a source of many insights into the three-dimensional conformation of proteins of known sequence but unknown structure.

3.6.5 Protein Modification and Cleavage Confer New Capabilities

Proteins are able to perform numerous functions relying solely on the versatility of their 20 amino acids. However, many proteins are covalently modifed, through the attachment of groups other than amino acids, to augment their functions (Figure 3.59). For example, *acetyl groups* are attached to the amino termini of many proteins, a modification that

```
 200 ?T(\G{+s x[A.N5~,#ATxSGpn`e□@
 400 oDr'Jh7s DFR:W4l'u+^v6zpJseOi
 600 e2ih'8zs n527x8l8d_ih=Hldseb.
 800 S#dh>}/s ]tZqC%lP%DK<|!^aseZ.
1000 VOth>nLs ut/isjl_kwojjwMasef.
1200 juth+nvs it is[lukh?SCw=ase5.
1400 Iithdn4s it isOl/ks/IxWLase~.
1600 M?thinrs it is lXk?T"_woasel.
1800 MSthinWs it is lwkN7□Kw(asel.
2000 Mhthin`s it is likv,aww_asel.
2200 MMthinns it is lik+5avwlasel.
2400 MethinXs it is likydaqw)asel.
2600 Methin4s it is lik2daseweasel.
2800 MethinHs it is like□aTweasel.
2883 Methinks it is like a weasel.

 200 )z~hg)W4{{cu!kO{d6jS!NlEyUx}p
 400 "W hi\kR.<&CfA%4-YlG!iT$6({|6
 600 .L=hinkm4(uMGP^lAWoE6klwW=yiS
 800  AthinkaPa_vYH liR\Hb,Uo4\-"(
1000 OFthinksP)@fZO li8v] /+Eln26B
1200 6ithinksMVt -V likm+g1#K~}BFk
1400 vxthinksaEt □w like.SlGeutks.
1600 :Othinks<it MC likesN2[eaVe4.
1800 uxthinksqit 0r likeQh)weaoeW.
2000 Y/thinks it id like7alwea)e&.
2200 Methinks it iW like a[weaWel.
2400 Methinks it is like a;weasel.
2431 Methinks it is like a weasel.
```

FIGURE 3.58 Typing monkey anology. A monkey randomly poking a typewriter could write a line from Shakespeare's *Hamlet,* provided that correct keystrokes were retained. In the two computer simulations shown, the cumulative number of keystrokes is given at the left of each line.

Novo Nordisk Ex. 2076, P. 31
Mylan Institutional v. Novo Nordisk
IPR2020-00324

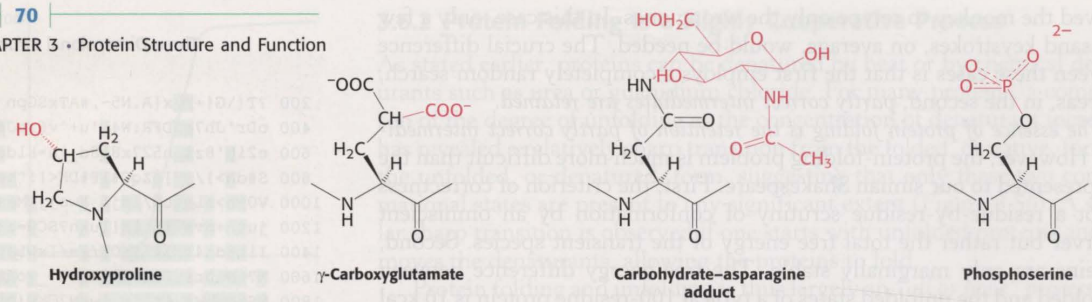


FIGURE 3.59 Finishing touches.
Some common and important covalent modifications of amino acid side chains are shown.

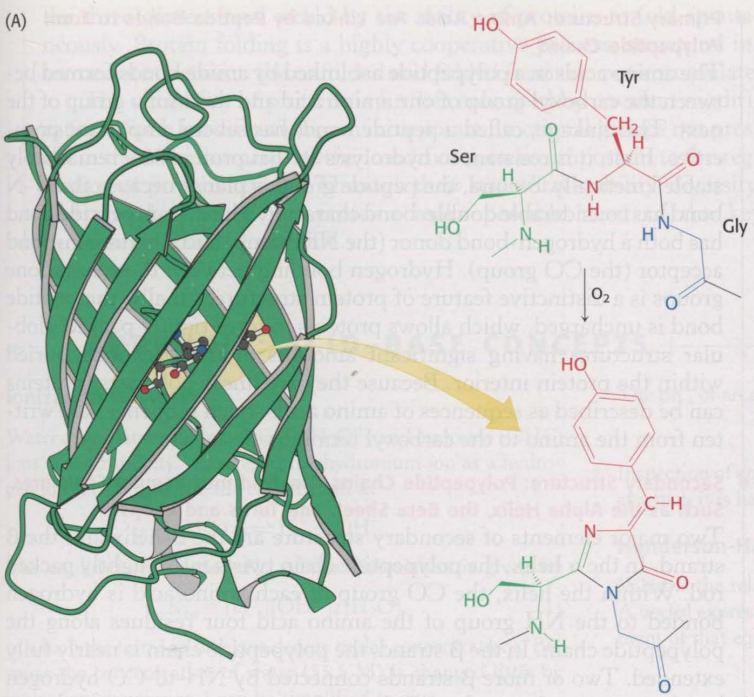
makes these proteins more resistant to degradation. The addition of *hydroxyl groups* to many proline residues stabilizes fibers of newly synthesized collagen, a fibrous protein found in connective tissue and bone. The biological significance of this modification is evident in the disease scurvy: a deficiency of vitamin C results in insufficient hydroxylation of collagen and the abnormal collagen fibers that result are unable to maintain normal tissue strength. Another specialized amino acid produced by a finishing touch is *γ-carboxyglutamate*. In vitamin K deficiency, insufficient carboxylation of glutamate in prothrombin, a clotting protein, can lead to hemorrhage. Many proteins, especially those that are present on the surfaces of cells or are secreted, acquire *carbohydrate units* on specific asparagine residues. The addition of sugars makes the proteins more hydrophilic and able to participate in interactions with other proteins. Conversely, the addition of a *fatty acid* to an α-amino group or a cysteine sulfhydryl group produces a more hydrophobic protein.

Many hormones, such as epinephrine (adrenaline), alter the activities of enzymes by stimulating the phosphorylation of the hydroxyl amino acids serine and threonine; *phosphoserine* and *phosphothreonine* are the most ubiquitous modified amino acids in proteins. Growth factors such as insulin act by triggering the phosphorylation of the hydroxyl group of tyrosine residues to form *phosphotyrosine*. The phosphoryl groups on these three modified amino acids are readily removed; thus they are able to act as reversible switches in regulating cellular processes. The roles of phosphorylation in signal transduction will be discussed extensively in Chapter 15.

The preceding modifications consist of the addition of special groups to amino acids. Other special groups are generated by chemical rearrangements of side chains and, sometimes, the peptide backbone. For example, certain jellyfish produce a fluorescent green protein (Figure 3.60). The source of the fluorescence is a group formed by the spontaneous rearrangement and oxidation of the sequence Ser-Tyr-Gly within the center of the protein. This protein is of great utility to researchers as a marker within cells (Section 4.3.5).

Finally, many proteins are cleaved and trimmed after synthesis. For example, digestive enzymes are synthesized as inactive precursors that can be stored safely in the pancreas. After release into the intestine, these precursors become activated by peptide-bond cleavage. In blood clotting, peptide-bond cleavage converts soluble fibrinogen into insoluble fibrin. A number of polypeptide hormones, such as adrenocorticotropic hormone, arise from the splitting of a single large precursor protein. Likewise, many virus proteins are produced by the cleavage of large polyprotein precursors. We shall encounter many more examples of modification and cleavage as essential features of protein formation and function. Indeed, these finishing touches account for much of the versatility, precision, and elegance of protein action and regulation.

Novo Nordisk Ex. 2076, P. 32
Mylan Institutional v. Novo Nordisk
IPR2020-00324



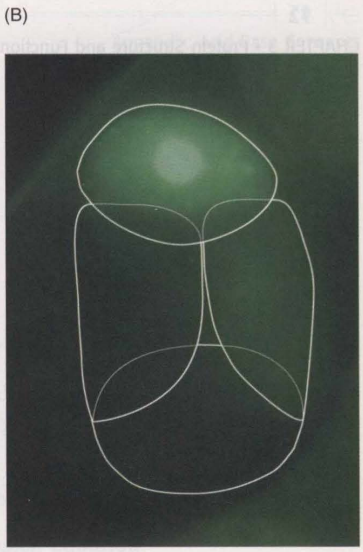


FIGURE 3.60 Chemical rearrangement in GFP. (A) The structure of green fluorescent protein (GFP). The rearrangement and oxidation of the sequence Ser-Try-Gly is the source of fluorescence. (B) Fluorescence micrograph of a four-cell embryo (cells are outlined) from the roundworm *C. elegans* containing a protein, PIE-1, labeled with GFP. The protein is expressed only in the cell (top) that will give rise to the germline. [(B) Courtesy of Geraldine Seydoux.]

SUMMARY

- Proteins are the workhorses of biochemistry, participating in essentially all cellular processes. Protein structure can be described at four levels. The primary structure refers to the amino acid sequence. The secondary structure refers to the conformation adopted by local regions of the polypeptide chain. Tertiary structure describes the overall folding of the polypeptide chain. Finally, quaternary structure refers to the specific association of multiple polypeptide chains to form multisubunit complexes.

- **Proteins Are Built from a Repertoire of 20 Amino Acids**
  Proteins are linear polymers of amino acids. Each amino acid consists of a central tetrahedral carbon atom linked to an amino group, a carboxylic acid group, a distinctive side chain, and a hydrogen. These tetrahedral centers, with the exception of that of glycine, are chiral; only the L isomer exists in natural proteins. All natural proteins are constructed from the same set of 20 amino acids. The side chains of these 20 building blocks vary tremendously in size, shape, and the presence of functional groups. They can be grouped as follows: (1) aliphatic side chains—glycine, alanine, valine, leucine, isoleucine, methionine, and proline; (2) aromatic side chains—phenylalanine, tyrosine, and tryptophan; (3) hydroxyl-containing aliphatic side chains—serine and threonine; (4) sulfhydryl-containing cysteine; (5) basic side chains—lysine, arginine, and histidine; (6) acidic side chains—aspartic acid and glutamic acid; and (7) carboxamide-containing side chains—asparagine and glutamine. These groupings are somewhat arbitrary and many other sensible groupings are possible.

Novo Nordisk Ex. 2076, P. 33
Mylan Institutional v. Novo Nordisk
IPR2020-00324

Primary Structure: Amino Acids Are Linked by Peptide Bonds to Form Polypeptide Chains

The amino acids in a polypeptide are linked by amide bonds formed between the carboxyl group of one amino acid and the amino group of the next. This linkage, called a peptide bond, has several important properties. First, it is resistant to hydrolysis so that proteins are remarkably stable kinetically. Second, the peptide group is planar because the C–N bond has considerable double-bond character. Third, each peptide bond has both a hydrogen-bond donor (the NH group) and a hydrogen-bond acceptor (the CO group). Hydrogen bonding between these backbone groups is a distinctive feature of protein structure. Finally, the peptide bond is uncharged, which allows proteins to form tightly packed globular structures having significant amounts of the backbone buried within the protein interior. Because they are linear polymers, proteins can be described as sequences of amino acids. Such sequences are written from the amino to the carboxyl terminus.

Secondary Structure: Polypeptide Chains Can Fold into Regular Structures Such as the Alpha Helix, the Beta Sheet, and Turns and Loops

Two major elements of secondary structure are the α helix and the β strand. In the α helix, the polypeptide chain twists into a tightly packed rod. Within the helix, the CO group of each amino acid is hydrogen bonded to the NH group of the amino acid four residues along the polypeptide chain. In the β strand, the polypeptide chain is nearly fully extended. Two or more β strands connected by NH-to-CO hydrogen bonds come together to form β sheets.

Tertiary Structure: Water-Soluble Proteins Fold into Compact Structures with Nonpolar Cores

The compact, asymmetric structure that individual polypeptides attain is called tertiary structure. The tertiary structures of water-soluble proteins have features in common: (1) an interior formed of amino acids with hydrophobic side chains and (2) a surface formed largely of hydrophilic amino acids that interact with the aqueous environment. The driving force for the formation of the tertiary structure of water-soluble proteins is the hydrophobic interactions between the interior residues. Some proteins that exist in a hydrophobic environment, in membranes, display the inverse distribution of hydrophobic and hydrophilic amino acids. In these proteins, the hydrophobic amino acids are on the surface to interact with the environment, whereas the hydrophilic groups are shielded from the environment in the interior of the protein.

Quaternary Structure: Polypeptide Chains Can Assemble into Multisubunit Structures

Proteins consisting of more than one polypeptide chain display quaternary structure, and each individual polypeptide chain is called a subunit. Quaternary structure can be as simple as two identical subunits or as complex as dozens of different subunits. In most cases, the subunits are held together by noncovalent bonds.

The Amino Acid Sequence of a Protein Determines Its Three-Dimensional Structure

The amino acid sequence completely determines the three-dimensional structure and, hence, all other properties of a protein. Some proteins can be unfolded completely yet refold efficiently when placed under conditions in which the folded form of the protein is stable. The amino acid sequence of a protein is determined by the sequences of bases in a DNA molecule. This one-dimensional sequence information is extended into

Novo Nordisk Ex. 2076, P. 34
Mylan Institutional v. Novo Nordisk
IPR2020-00324

the three-dimensional world by the ability of proteins to fold spontaneously. Protein folding is a highly cooperative process; structural intermediates between the unfolded and folded forms do not accumulate.

The versatility of proteins is further enhanced by covalent modifications. Such modifications can incorporate functional groups not present in the 20 amino acids. Other modifications are important to the regulation of protein activity. Through their structural stability, diversity, and chemical reactivity, proteins make possible most of the key processes associated with life.

APPENDIX: ACID–BASE CONCEPTS

Ionization of Water

Water dissociates into hydronium (H_3O^+) and hydroxyl (OH^-) ions. For simplicity, we refer to the hydronium ion as a hydrogen ion (H^+) and write the equilibrium as

$$H_2O \rightleftharpoons H^+ + OH^-$$

The equilibrium constant K_{eq} of this dissociation is given by

$$K_{eq} = [H^+][OH^-]/[H_2O] \quad (1)$$

in which the terms in brackets denote molar concentrations. Because the concentration of water (55.5 M) is changed little by ionization, expression 1 can be simplified to give

$$K_W = [H^+][OH^-] \quad (2)$$

in which K_w is the ion product of water. At 25°C, K_w is 1.0×10^{-14}.

Note that the concentrations of H^+ and OH^- are reciprocally related. If the concentration of H^+ is high, then the concentration of OH^- must be low, and vice versa. For example, if $[H^+] = 10^{-2}$ M, then $[OH^-] = 10^{-12}$ M.

Definition of Acid and Base

An acid is a proton donor. A base is a proton acceptor.

$$\text{Acid} \rightleftharpoons H^+ + \text{base}$$

$$\underset{\text{Acetic acid}}{CH_3COOH} \rightleftharpoons H^+ + \underset{\text{Acetate}}{CH_3COO^-}$$

$$\underset{\text{Ammonium ion}}{NH_4^+} \rightleftharpoons H^+ + \underset{\text{Ammonia}}{NH_3}$$

The species formed by the ionization of an acid is its conjugate base. Conversely, protonation of a base yields its conjugate acid. Acetic acid and acetate ion are a conjugate acid–base pair.

Definition of pH and p*K*

The pH of a solution is a measure of its concentration of H^+. The pH is defined as

$$pH = \log_{10}(1/[H^+]) = -\log_{10}[H^+] \quad (3)$$

The ionization equilibrium of a weak acid is given by

$$HA \rightleftharpoons H^+ + A^-$$

The apparent equilibrium constant K_a for this ionization is

$$K_a = [H^+][A^-]/[HA] \quad (4)$$

The pK_a of an acid is defined as

$$pK_a = -\log K_a = \log(1/K_a) \quad (5)$$

Inspection of equation 4 shows that the pK_a of an acid is the pH at which it is half dissociated, when $[A^-] = [HA]$.

Henderson-Hasselbalch Equation

What is the relation between pH and the ratio of acid to base? A useful expression can be derived from equation 4. Rearrangement of that equation gives

$$1/[H^+] = 1/K_a[A^-]/[HA] \quad (6)$$

Taking the logarithm of both sides of equation 6 gives

$$\log(1/[H^+]) = \log(1/K_a) + \log([A^-]/[HA]) \quad (7)$$

Substituting pH for $\log 1/[H^+]$ and pK_a for $\log 1/K_a$ in equation 7 yields

$$pH = pK_a + \log([A^-]/[HA]) \quad (8)$$

which is commonly known as the Henderson-Hasselbalch equation.

The pH of a solution can be calculated from equation 8 if the molar proportion of A^- to HA and the pK_a of HA are known. Consider a solution of 0.1 M acetic acid and 0.2 M acetate ion. The pK_a of acetic acid is 4.8. Hence, the pH of the solution is given by

$$pH = 4.8 + \log(0.2/0.1) = 4.8 + \log 2.0 = 4.8 + 0.3 = 5.1$$

Conversely, the pK_a of an acid can be calculated if the molar proportion of A^- to HA and the pH of the solution are known.

Buffers

An acid–base conjugate pair (such as acetic acid and acetate ion) has an important property: it resists changes in the pH of a solution. In other words, it acts as a *buffer*. Consider the addition of OH^- to a solution of acetic acid (HA):

$$HA + OH^- \rightleftharpoons A^- + H_2O$$

A plot of the dependence of the pH of this solution on the amount of OH^- added is called a *titration curve* (Figure 3.61). Note that there is an inflection point in the curve at pH 4.8, which is the pK_a of acetic acid. In the vicinity of this pH, a relatively large amount of OH^- produces little change in pH. In other words, the buffer maintains the value of pH near a given value, despite the addition of other either protons or hydroxide

Novo Nordisk Ex. 2076, P. 35
Mylan Institutional v. Novo Nordisk
IPR2020-00324

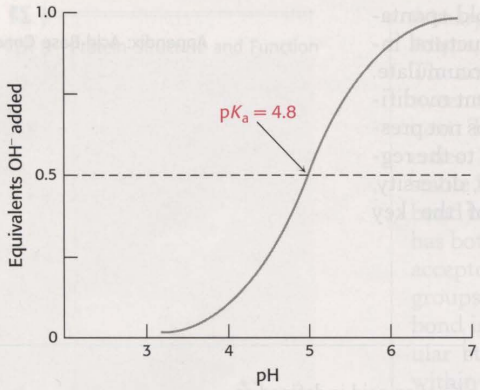


FIGURE 3.61 Titration curve of acetic acid.

ions. In general, a weak acid is most effective in buffering against pH changes in the vicinity of its pK_a value.

pK_a Values of Amino Acids

An amino acid such as glycine contains two ionizable groups: an α-carboxyl group and a protonated α-amino group. As base is added, these two groups are titrated (Figure 3.62). The pK_a of the α-COOH group is 2.4, whereas that of the α-NH_3^+ group is 9.8. The pK_a values of these groups in other amino acids are similar (Table 3.4). Some amino acids, such as aspartic acid, also contain an ionizable side chain. The pK_a values of ionizable side chains in amino acids range from 3.9 (aspartic acid) to 12.5 (arginine).

TABLE 3.4 pK_a values of some amino acids

Amino acid	pK_a values (25°C) α-COOH group	α-NH_3^+ group	Side chain
Alanine	2.3	9.9	
Glycine	2.4	9.8	
Phenylalanine	1.8	9.1	
Serine	2.1	9.2	
Valine	2.3	9.6	
Aspartic acid	2.0	10.0	3.9
Glutamic acid	2.2	9.7	4.3
Histidine	1.8	9.2	6.0
Cysteine	1.8	10.8	8.3
Tyrosine	2.2	9.1	10.9
Lysine	2.2	9.2	10.8
Arginine	1.8	9.0	12.5

After J. T. Edsall and J. Wyman, *Biophysical Chemistry* (Academic Press, 1958), Chapter 8.

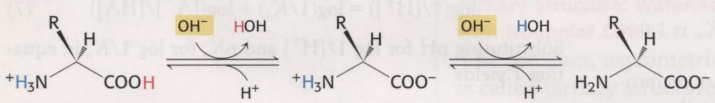


FIGURE 3.62 Titration of the α-carboxyl and α-amino groups of an amino acid.

KEY TERMS

side chain (R group) (p. 43)
L amino acid (p. 43)
dipolar ion (zwitterion) (p. 43)
peptide bond (amide bond) (p. 51)
disulfide bond (p. 52)
primary structure (p. 53)
phi (ϕ) angle (p. 55)
psi (ψ) angle (p. 55)
Ramachandran diagram (p. 55)
α helix (p. 56)
β pleated sheet (p. 58)
β strand (p. 58)
reverse turn (β turn; hairpin turn) (p. 60)
secondary structure (p. 61)
tertiary structure (p. 62)
domain (p. 63)
subunit (p. 64)
quaternary structure (p. 64)
cooperative transition (p. 68)

SELECTED READINGS

Where to start

Richardson, J. S., 1981.The anatomy and taxonomy of protein structure. *Adv. Protein Chem.* 34:167–339.
Doolittle, R. F., 1985. Proteins. *Sci. Am.* 253(4):88–99.
Richards, F. M., 1991. The protein folding problem. *Sci. Am.* 264(1): 54–57.
Weber, A. L., and Miller, S. L., 1981. Reasons for the occurrence of the twenty coded protein amino acids. *J. Mol. Evol.* 17:273–284.

Books

Branden, C., Tooze, J., 1999. *Introduction to Protein Structure* (2d ed.). Garland.
Perutz, M. F., 1992. *Protein Structure: New Approaches to Disease and Therapy.* W. H. Freeman and Company.
Creighton, T. E., 1992. *Proteins: Structures and Molecular Principles* (2d ed.). W. H. Freeman and Company.
Schultz, G. E., and Schirmer, R. H., 1979. *Principles of Protein Structure.* Springer-Verlag.

Conformation of proteins

Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., Donlan, M. E., and Suries, M. C., 1992. Looking at proteins: Representations, folding, packing, and design. *Biophys. J.* 63: 1186–1220.

Novo Nordisk Ex. 2076, P. 36
Mylan Institutional v. Novo Nordisk
IPR2020-00324

Chothia, C., and Finkelstein, A. V., 1990. The classification and origin of protein folding patterns. *Annu. Rev. Biochem.* 59:1007–1039.

Alpha helices, beta sheets, and loops

O'Neil, K. T., and DeGrado, W. F., 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* 250:646–651.

Zhang, C., and Kim, S. H., 2000. The anatomy of protein beta-sheet topology. *J. Mol. Biol.* 299:1075–1089.

Regan, L., 1994. Protein structure: Born to be beta. *Curr. Biol.* 4:656–658.

Leszczynski, J. F., and Rose, G. D., 1986. Loops in globular proteins: A novel category of secondary structure. *Science* 234:849–855.

Srinivasan, R., and Rose, G. D., 1999. A physical basis for protein secondary structure. *Proc. Natl. Acad. Sci. USA* 96:14258–14263.

Domains

Bennett, M. J., Choe, S., and Eisenberg, D., 1994. Domain swapping: Entangling alliances between proteins. *Proc. Natl. Acad. Sci. USA* 91:3127–3131.

Bergdoll, M., Eltis, L. D., Cameron, A. D., Dumas, P., and Bolin, J. T., 1998. All in the family: Structural and evolutionary relationships among three modular proteins with diverse functions and variable assembly. *Protein Sci.* 7:1661–1670.

Hopfner, K. P., Kopetzki, E., Kresse, G. B., Bode, W., Huber, R., and Engh, R. A., 1998. New enzyme lineages by subdomain shuffling. *Proc. Natl. Acad. Sci. USA* 95:9813–9818.

Ponting, C. P., Schultz, J., Copley, R. R., Andrade, M. A., and Bork, P., 2000. Evolution of domain families. *Adv. Protein Chem.* 54: 185–244.

Protein folding

Anfinsen, C. B., 1973. Principles that govern the folding of protein chains. *Science* 181:223–230.

Baldwin, R. L., and Rose, G. D., 1999. Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* 24:26–33.

Baldwin, R. L., and Rose, G. D., 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* 24:77–83.

Staley, J. P., and Kim, P. S., 1990. Role of a subdomain in the folding of bovine pancreatic trypsin inhibitor. *Nature* 344:685–688.

Neira, J. L., and Fersht, A. R., 1999. Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments. *J. Mol. Biol.* 285:1309–1333.

Covalent modification of proteins

Krishna, R. G., and Wold, F., 1993. Post-translational modification of proteins. *Adv. Enzymol. Relat. Areas. Mol. Biol.* 67:265–298.

Aletta, J. M., Cimato, T. R., and Ettinger, M. J., 1998. Protein methylation: A signal event in post-translational modification. *Trends Biochem. Sci.* 23:89–91.

Glazer, A. N., DeLange, R. J., and Sigman, D. S., 1975. *Chemical Modification of Proteins.* North-Holland.

Tsien, R. Y., 1998. The green fluorescent protein. *Annu. Rev. Biochem.* 67:509–544.

Molecular graphics

Kraulis, P., 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* 24:946–950.

Ferrin, T., Huang, C., Jarvis, L., and Langridge, R., 1988. The MIDAS display system. *J. Mol. Graphics* 6:13–27.

Richardson, D. C., and Richardson, J. S., 1994. Kinemages: Simple macromolecular graphics for interactive teaching and publication. *Trends Biochem. Sci.* 19:135–138.

PROBLEMS

1. *Shape and dimension.* (a) Tropomyosin, a 70-kd muscle protein, is a two-stranded α-helical coiled coil. Estimate the length of the molecule? (b) Suppose that a 40-residue segment of a protein folds into a two-stranded antiparallel β structure with a 4-residue hairpin turn. What is the longest dimension of this motif?

2. *Contrasting isomers.* Poly-L-leucine in an organic solvent such as dioxane is α helical, whereas poly-L-isoleucine is not. Why do these amino acids with the same number and kinds of atoms have different helix-forming tendencies?

3. *Active again.* A mutation that changes an alanine residue in the interior of a protein to valine is found to lead to a loss of activity. However, activity is regained when a second mutation at a different position changes an isoleucine residue to glycine. How might this second mutation lead to a restoration of activity?

4. *Shuffle test.* An enzyme that catalyzes disulfide–sulfhydryl exchange reactions, called protein disulfide isomerase (PDI), has been isolated. PDI rapidly converts inactive scrambled ribonuclease into enzymatically active ribonuclease. In contrast, insulin is rapidly inactivated by PDI. What does this important observation imply about the relation between the amino acid sequence of insulin and its three-dimensional structure?

5. *Stretching a target.* A protease is an enzyme that catalyzes the hydrolysis of the peptide bonds of target proteins. How might a protease bind a target protein so that its main chain becomes fully extended in the vicinity of the vulnerable peptide bond?

6. *Often irreplaceable.* Glycine is a highly conserved amino acid residue in the evolution of proteins. Why?

7. *Potential partners.* Identify the groups in a protein that can form hydrogen bonds or electrostatic bonds with an arginine side chain at pH 7.

8. *Permanent waves.* The shape of hair is determined in part by the pattern of disulfide bonds in keratin, its major protein. How can curls be induced?

9. *Location is everything.* Proteins that span biological membranes often contain α helices. Given that the insides of membranes are highly hydrophobic (Section 12.2.1), predict what type of amino acids would be in such a helix. Why is an α helix particularly suited to exist in the hydrophobic environment of the interior of a membrane?

10. *Issues of stability.* Proteins are quite stable. The lifetime of a peptide bond in aqueous solution is nearly 1000 years. However, the $\Delta G^{\circ\prime}$ of hydrolysis of proteins is negative and quite large. How can you account for the stability of the peptide bond in light of the fact that hydrolysis releases much energy?

11. *Minor species.* For an amino acid such as alanine, the major species in solution at pH 7 is the zwitterionic form. Assume a pK_a value of 8 for the amino group and a pK_a value of 3 for the

Novo Nordisk Ex. 2076, P. 37
Mylan Institutional v. Novo Nordisk
IPR2020-00324

carboxylic acid and estimate the ratio of the concentration of neutral amino acid species (with the carboxylic acid protonated and the amino group neutral) to that of the zwitterionic species at pH 7.

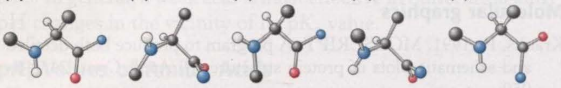
12. *A matter of convention.* All L amino acids have an *S* absolute configuration except L-cysteine, which has the *R* configuration. Explain why L-cysteine is designated as the *R* absolute configuration.

13. *Hidden message.* Translate the following amino acid sequence into one-letter code: Leu-Glu-Ala-Arg-Asn-Ile-Asn-Gly-Ser-Cys-Ile-Glu-Cys-Glu-Ile-Ser-Gly-Arg-Glu-Ala-Thr.

14. *Who goes first?* Would you expect Pro–X peptide bonds to tend to have cis conformations like those of X–Pro bonds? Why or why not?

15. *Matching.* For each of the amino acid derivatives shown below (A–E), find the matching set of ϕ and ψ values (a–e).

(A) (B) (C) (D) (E)



(a)	(b)	(c)	(d)	(e)
$\phi = 120°$, $\psi = 120°$	$\phi = 180°$, $\psi = 0°$	$\phi = 180°$, $\psi = 180°$	$\phi = 0°$, $\psi = 180°$	$\phi = -60°$, $\psi = -40°$

16. *Concentrate of the concentration.* A solution of a protein whose sequence includes three tryptophan residues, no tyrosine residues, and no phenylalanine residues has an absorbance of 0.1 at 280 nm in a cell with a path length of 1 cm. Estimate the concentration of the protein in units of molarity. If the protein has a molecular mass of 100 kd, estimate the concentration in units of milligrams of protein per milliliter of solution.

Media Problem

You can use the **Structural Insights** and **Conceptual Insights** as visual aids to help you answer Media Problems. Go to the Website: www.whfreeman.com/biochem5, and select the applicable module.

17. *Inside-out, back-to-front.* In the Media Problem section of the **Structural Insights** module on protein structure, you can examine molecular models of four putative protein structures. One of the four structures has been determined by x-ray crystallography. The other three have been made-up, and in fact are very unlikely to occur. Which are the structures that are unlikely to occur and why?

Novo Nordisk Ex. 2076, P. 38
Mylan Institutional v. Novo Nordisk
IPR2020-00324